EL2800 Reinforcement Learning

Re-exam – April 2019

Department of Automatic Control
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology

Re-exam (Omtenta), April 17, 2019, kl 8.00 - 13.00

Aids. Slides of the lectures (**not exercises**), blackboard notes, mathematical tables.

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.

The exam consists in 5 problems. The distribution of points among these problems is indicated below.

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 May 1, 2019

Good luck!

Problem 1

Provide short answers to the following questions **and** a short motivation (not more than ca 5 sentences per question).

a) What is the Markov property? [1 pts]

b) What is *experience replay* and why is it useful? [1 pts]

c) Suppose x_k is a sequence of random variables that fullfills

$$\Pr\{x_k | x_{k-1}, x_{k-2}, x_{k-3}, \dots\} = \Pr\{x_k | x_{k-1}, x_{k-2}\}.$$

Let $z_k = (x_k, x_{k-1})$. Is z_k a Markov Chain? [1 pts]

d) Name three persons that have made important contributions to RL and/or MDPs, and summarize each one's contribution in one sentence. [1 pts]

e) Provide two interpretations of the discount factor λ in an MDP with infinite horizon objective $\mathbb{E}\{\sum_{t=0}^{\infty} \lambda^t r(s_t, a_t)\}$. [1 pts]

f) Mention a setting in which one cannot employ Monte Carlo methods, and TD methods have to be employed. Motivate. [1 pts]

g) Consider a discounted RL problem with discount factor λ. Give a minimax lower bound on the sample complexity of a RL algorithm to determine an ϵ-optimal policy with probability $1 - \delta$. Why do we need to resort to function approximation when the state space is large? [2 pts]

h) Assume that we run a ϵ-greedy policy in SARSA algorithm, but ϵ varies over time and is equal to $1/t$ at step t. What is the average number times we explore random actions until time T? [1 pts]

i) Suppose we take the step-size $\alpha_k = 1/\sqrt{k}$ in the Q-learning algorithm. Are the iterates guaranteed to converge almost surely to the true Q-function? [1 pts]

Solution:

a) $\Pr\{x_k|x_{k-1},,\ldots,x_0\} = \Pr\{x_k|x_{k-1}\}$

b) In Q-learning with function approximation, successive updates are strongly correlated (since they follow a particular trajectory). In order to improve the convergence rate, with experience replay, one maintains a buffer B of previous experiences (s,a,r,s') and then samples mini-batches of fixed size k from B uniformly at random.

c) Yes;

$$\Pr\{z_k|z_{k-1},z_{k-2},z_{k-3},\ldots\} = \Pr\{x_k,x_{k-1}|x_{k-1},x_{k-2},x_{k-3},x_{k-4},\ldots\}$$

From the given property, x_k depends only on x_{k-1} and x_{k-2}, The variable x_{k-1} depends on x_{k-2} and x_{k-3}, but since its value is given (we condition on x_{k-1}), x_{k-2} and x_{k-3} provide no additional information. Hence, we can remove everything except x_{k-1} and x_{k-2} in the conditioning.

$$\begin{aligned} &= \Pr\{x_k,x_{k-1}|x_{k-1},x_{k-2}\} \\ &= \Pr\{z_k|z_{k-1}\}. \end{aligned}$$

d)
- Richard Bellman: Inventor of dynamic programming.
- Christopher Watkins: Inventor of Q-learning.
- Herbert Robbins and Sutton Monro: Formulated the stochastic approximation algorithm.
- Ronald Williams: Inventor of the REINFORCE algorithm.
- etc.

e) *Interest rate:* The value of a unit reward decreases with time at geometric rate λ. *Random time horizon:* the decision maker has a time horizon T that is geometrically distributed.

f) When the episode does not have finite length.

g) The lower bound is $\frac{SA}{\epsilon^2(1-\lambda)^3}\log(\delta^{-1})$. When SA is large, we cannot treat RL problems without approximations. Function approximation allows us to remove the term SA in the sample complexity.

h) We have $\mathbb{P}[E_t] = 1/t$ where E_t is the event that a random action is explored at step t. Hence, the average number of times we explore is:

$$1 + \frac{1}{2} + \ldots + \frac{1}{T} = \ln T + \gamma + O(1/T),$$

where γ is the Euler-Mascheroni constant.

i) No. There are two criteria that have to be fulfilled by the step-sizes α_k: $\sum \alpha_k = \infty$ and $\sum \alpha_k^2 < \infty$. We have that

$$\sum_k \alpha_k = \sum_k \frac{1}{\sqrt{k}} = \infty, \quad \textbf{Ok!}$$

but

$$\sum_k \alpha_k^2 = \sum_k \frac{1}{k} = \infty. \quad \textbf{Not Ok!}$$

Problem 2

Let Y_k be an independent and identically distributed (i.i.d.) sequence of random variables taking values in the set $\{0, 1, 2\}$ according to the probability mass function:

$$\Pr\{Y_k = i\} = \begin{cases} p_0 & \text{if } i = 0, \\ p_1 & \text{if } i = 1, \\ p_2 & \text{if } i = 2, \end{cases}$$

with $p_0, p_1, p_2 \geq 0$ and $p_0 + p_1 + p_2 = 1$. Consider the Markov chain X_k defined via[1]

$$X_k = X_{k-1} + Y_k \pmod 3,$$

with $X_0 = 0$.

a) What is the state-space of the Markov chain X_k? [1 pt]

b) What is its transition matrix? [1 pt]

c) Show that X_k's stationary distribution is the uniform distribution. Can you identify the property of the transition matrix that guarantees that the uniform distribution is the stationary distribution? [2 pts]

d) Assume that $p_1 = 0$ and $p_0 = p_2 = 1/2$. Specify the communicating classes, and determine which ones are recurrent and which ones are transient. [1 pt]

Model, if this is at all possible, the two following problems using a Markov Decision Process. Precise the time horizon, state and action spaces, the transition probabilities, and the rewards. *Do not try to solve these MDPs.*

e) A manuscript must be submitted in N days and has been typed with a known number of mistakes M. Mistakes may be found and corrected through a review. Each review takes one day to complete and costs an amount $c_1 > 0$. On the k^{th} review, each undetected mistake is found independently with probability p_k. Each undetected mistake left in the manuscript when it is sent to the printer costs an amount $c_2 > 0$. The problem is to decide when to stop reviewing and send the manuscript to the printer. [5 pts]

[1] Recall that the modulo operation finds the remainder after division of one number by another. For example, 11 (mod 10) = 1 and 6 (mod 20) = 6.

Solution:

a) $\{0, 1, 2\}$

b)

$$P = \begin{bmatrix} p_0 & p_1 & p_2 \\ p_2 & p_0 & p_1 \\ p_1 & p_2 & p_0 \end{bmatrix}$$

b) We need to check that $\pi_{\mathrm{u}} = (1/3, 1/3, 1/3)$ satisfies $\pi_u P = \pi_u$:

$$\begin{aligned} \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix} \begin{bmatrix} p_0 & p_1 & p_2 \\ p_2 & p_0 & p_1 \\ p_1 & p_2 & p_0 \end{bmatrix} &= \begin{bmatrix} \frac{1}{3}(p_0 + p_2 + p_1) & \frac{1}{3}(p_1 + p_0 + p_2) & \frac{1}{3}(p_2 + p_1 + p_0) \end{bmatrix} \\ &= \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix}. \end{aligned}$$

This holds for all *doubly stochastic matrices* (i.e., those matrices where both the elements of the rows and the columns sum to one).

d) Communicating class: $\{0, 1, 2\}$. Transient: $\emptyset$. Recurrent: $\{0, 1, 2\}$.

e) The state space is $\mathcal{S} = (\{0, \ldots, N\} \times \{0, \ldots, M\}) \cup \{\mathit{printed}\}$ such that the state s is either the terminal state *printed* or a tuple (n, m) where n is the number of proofreadings carried out so far and m is the remaining number of mistakes in the manuscript. The action space is $A = \{\mathit{review}, \mathit{print}\}$. We have an MDP with finite horizon N and rewards: $r((n, m), \mathit{review}) = -c_1$, $r((n, m), \mathit{print}) = -c_2 m$, and $r(\mathit{printed}, \cdot) = 0$. The non-zero transition probabilities are $p((n + 1, m - j) | (n, m), \mathit{review}) = \binom{m}{j} p_{n+1}^j (1 - p_{n+1})^{m-j}$ for $j = \{0, \ldots, m\}$, $p(\mathit{printed} | (n, m), \mathit{print}) = 1$, and $p(\mathit{printed} | \mathit{printed}, \cdot) = 1$.

Problem 3

You wish to sell your house in Los Angeles. You try to sell it every spring, and start year 1. Due to climate changes, the risk of your house to burn is increasing summer after summer. In year t, the probability that your house disappears due to wildfires is b_t. Each spring you receive offers whose maximum is i.i.d. (across years) and with distribution described by $f(w)$, the probability that the best offer is w, for $w \in \{1, \dots, W\}$. After selling your house, you place the money and enjoy an interest rate of $r\%$. Your objective is to maximize the average amount of money at the end of year $T > 1$.

a) Model this problem as an MDP (describe the MDP in full detail). [3 pts]

b) Establish that the optimal policy is threshold-based, i.e., you decide to accept the best offer made in year t if this offer exceeds a threshold. [3 pts]

c) Provide a general recursive formula satisfied by the thresholds. [2 pts]

d) Now assume that the best offer distribution is uniform over $[0, 1]$ and that $b_t = b$ for all t. Answer the question c) again in this setting. When T is very large, what are the optimal decisions in the first years? [2 pts]

a) States $w \in \{0, 1, \dots, W\}$. If $w = 0$ it means your house has been sold or has burnt, otherwise, w is the best offer the current year.
Actions A (Accept) or R (Reject).
Rewards: year t, $r_t(w, A) = (1+r)^{T-t} w$, $r_t(w, R) = 0$, for all $w \in \{0, 1, \dots, W\}$.
Transitions: $p_t(w'|w, R) = (1 - b_t) f(w')$, $p_t(0|w, R) = b_t$, $p_t(0|w, A) = 1$.

b) Denote by $V_t(w)$ the value when in year t the state is w. Bellman's equation is: $V_T(w) = w$ (you have to accept the offer), and for $1 \le t < T$,

$$V_t(w) = 1_{\{w>0\}} \max\{(1+r)^{T-t} w, \sum_{w'} (1 - b_t) f(w') V_{t+1}(w')\}.$$

Hence it is optimal to accept the offer if and only if

$$w \ge \alpha_t := (1+r)^{t-T} \sum_{w'} (1 - b_t) f(w') V_{t+1}(w').$$

c) The thresholds satisfy: $\alpha_T = 0$, and for $t < T$,

$$\begin{aligned}
\alpha_t &= (1+r)^{t-N} \sum_{w'} (1 - b_t) f(w') V_{t+1}(w') \\
&= (1+r)^{t-N} \sum_{w'} (1 - b_t) f(w') \max\{(1+r)^{T-t-1} w', (1+r)^{T-t-1} \alpha_{t+1}\} \\
&= (1+r)^{-1} \sum_{w'} (1 - b_t) f(w') \max\{w', \alpha_{t+1}\}
\end{aligned}$$

d) We get:

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

When T is large, the first decisions are similar and correspond to a threshold, fixed point of the above recursion. We find:

$$\alpha_1 = \beta - \sqrt{\beta^2 - 1},$$

where $\beta = (1+r)/(1-b)$.

Problem 4

Consider a discounted MDP with $\mathcal{S} = \{\text{A, B, C}\}$ and $\mathcal{A} = \{a, b, c\}$. We plan to use either the Q-learning or the SARSA algorithm in order to learn to control the system. We initialize the estimated Q-function as all zeros – that is:

$$Q^{(0)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array} .$$

The observed trajectory is as follows (for these transitions, we are imposed a policy):

$$(?, ?, ?); (A, ?, ?); (B, a, 100); (A, b, 60); (B, c, 70); (C, b, 40); (A, a, 20); \ldots$$

where each triplet represents the state, the selected action, and the corresponding reward. Some of the information has been corrupted (marked with question marks) in the above sequence.

a) Before the information became corrupt, we ran the Q-learning algorithm and obtained that

$$Q^{(2)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 11 & 0 & 0 \\ 0 & 0 & 60 \\ 0 & 0 & 0 \end{bmatrix} \end{array} .$$

The discount factor was $\lambda = 0.5$ and the learning rate was fixed to $\alpha = 0.1$. Can you infer what the corrupt information was (i.e., the first state, the first and second selected actions, and the first and second observed rewards? [4 pts]

b) Provide the updated Q-values, using the Q-learning algorithm, after the whole sequence. Use the same values for λ and α as in a). [2 pts]

c) What is the current greedy policy after these updates? [1 pts]

d) Provide the updated Q-values after the whole sequence using the SARSA algorithm (initialized with $Q^{(0)}$ as all zeros). Take the first two (state, action, reward)-triplets as those given in your answer to a). Let the discount factor be $\lambda = 0.5$ and the learning rate fixed to $\alpha = 0.1$. [2 pts]

e) What is the current greedy policy after these updates? [1 pts]

Solution:

a) Since the entry (B,c) has been updated, it must correspond to the first state-action pair (since the second state is known to be A). We solve for the first reward R_1 in the Q-learning update rule as:

$$\begin{aligned}60 &= Q^{(1)}(B,c)\\ &= Q^{(0)}(B,c) + \alpha\left[R_1 + \lambda \max_{\phi} Q^{(0)}(A,\phi) - Q^{(0)}(B,c)\right]\\ &= 0 + 0.1 \times [R_1 + 0.5 \times 0 - 0]\,,\end{aligned}$$

which gives $R_1 = 600$. The second action must have been a (since the element (A,a) has been updated). Solving for the second reward R_2, as above, yields:

$$\begin{aligned}11 &= Q^{(2)}(A,a)\\ &= Q^{(1)}(A,a) + \alpha\left[R_2 + \lambda \max_{\phi} Q^{(1)}(B,\phi) - Q^{(1)}(A,a)\right]\\ &= 0 + 0.1 \times [R_2 + 0.5 \times 60 - 0]\,,\end{aligned}$$

which gives $R_2 = 110 - 30 = 80$. Hence, the full uncorrupted sequence was:

$$(B,c,600);(A,a,80);(B,a,100);(A,b,60);(B,c,70);(C,b,40);(A,a,20);\ldots$$

b)

$$Q^{(6)} = \begin{array}{c} \\ A \\ B \\ C \end{array}\!\!\begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 11 & 9 & 0 \\ 10.55 & 0 & 61 \\ 0 & 4.55 & 0 \end{bmatrix} \end{array}.$$

c) The greedy policy is $\pi(A) = a$, $\pi(B) = c$, $\pi(C) = b$.

d)

$$Q^{(6)} = \begin{array}{c} \\ A \\ B \\ C \end{array}\!\!\begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 8 & 9 & 0 \\ 10 & 0 & 61 \\ 0 & 4.4 & 0 \end{bmatrix} \end{array}.$$

e) The greedy policy is $\pi(A) = b$, $\pi(B) = c$, $\pi(C) = b$.

Problem 5

Policy gradient. We consider an episodic RL problem with finite state-space $\mathcal{S}$ and action space $\mathcal{A} = \{1, \dots, n+1\}$. For all states s, let $f(s)$ be a real valued function in $[1, 2]$. We parameterize the policy using parameter vector $\theta = (\theta_1, \dots, \theta_n) \in [0, 1]^n$ according to the following recursion: For $i \in \{1, \dots, n\}$, initialize $i = 1$ and draw independent random variable Z_i uniformly from $[0, f(s)]$. If $Z_i \leq \theta_i$, choose action $a = i$, otherwise, set $i \leftarrow i+1$ and repeat. At the last step of the recursion, if $Z_n > \theta_n$, choose $a = n + 1$.

a) Compute in state s, the probability $\pi_\theta(s, i)$ of choosing action i. [2 pts]

b) What is the Monte-Carlo REINFORCE update of θ upon observing an episode $\tau = (s_1, a_1, r_1, \dots, s_T, a_T, r_T)$? Provide explicit formulas using the function f, θ and τ only [3 pts]

On-policy control with function approximation. Consider a discounted RL problem, that we wish to solve using approximations of the (state, action) value function (i.e., parametrized by vector θ).

c) We observe the transition (s_t, a_t, r_t, s_{t+1}). State the Q update in the Q-learning algorithm with function approximation. Why is it a semi-gradient algorithm? [2 pts]

d) What is the difference between the targets of this algorithm compared to standard SGD and what are its implications? Propose a modification that addresses this problem. [3 pts]

a) We have:

$$\pi_\theta(s,1) = \frac{\theta}{f(s)}, \quad \pi_\theta(s,i) = \frac{\prod_{j=1}^{i-1}(f(s)-\theta_j)\,\theta_i}{f(s)^i} \text{ for } i \in \{2,\dots,n\}, \quad \pi_\theta(s,n{+}1) = \frac{\prod_{j=1}^{n}(f(s)-\theta_j)}{f(s)^n},$$

b) The algorithm is given slide 27 Part 5. We need to compute $\nabla_\theta \ln \pi_\theta(s,i)$ which will be of the form:

$$\nabla_\theta \ln \pi_\theta(s,i) = \left(\frac{\partial \ln \pi_\theta(s,i)}{\partial \theta_1}, \frac{\partial \ln \pi_\theta(s,i)}{\partial \ln \theta_2}, \dots, \frac{\partial \ln \pi_\theta(s,i)}{\partial \theta_i}, 0_{n-i}\right)^T$$

where 0_i is a zero row vector of length i. We will have:

$$\frac{\partial \ln \pi_\theta(s,i)}{\partial \theta_i} = \frac{1}{\theta_i}$$

and

$$\frac{\partial \ln \pi_\theta(s,i)}{\partial \theta_k} = \frac{1}{\theta_k - f(s)} \text{ for } k < i \text{ and } i \in \{2,\dots,n+1\}.$$

c) The Q-update is given in slide 25 Part 6 as:

$$\theta \leftarrow \theta + \alpha\left(r_t + \lambda \max_b Q_\theta(s_{t+1},b) - Q_\theta(s_t,a_t)\right)\nabla_\theta Q_\theta(s_t,a_t).$$

It is a semi-gradient algorithm because only half of the TD term is differentiated w.r.t. θ.

d) Unlike SGD, the target values change at every iteration due to their dependency on θ. If they change too quickly, it becomes impossible to converge, so a possible solution is to delay the update of the target by C number of iterations:

$$\theta \leftarrow \theta + \alpha\left(r_t + \lambda \max_b Q_\phi(s_{t+1},b) - Q_\theta(s_t,a_t)\right)\nabla_\theta Q_\theta(s_t,a_t).$$

Every C iterations, update $\phi \leftarrow \theta$.