EL2805 Reinforcement Learning

Exam – January 2019

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

Exam (tentamen), January 8, 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. Robert Matila 0762014056
Damianos Tranos 0732797826
Alexandre Proutiere 087906351

Results. Posted no later than January 21, 2019

Good luck!

Problem 1

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

a) Let $X_1, \ldots, X_k, \ldots$ be an homogenous Markov chain with values in a finite set $\mathcal{X}$ and transition probabilities $p(x, y) = \mathbb{P}[X_{k+1} = y|X_k = x]$ for all $(x, y) \in \mathcal{X}^2$. Then the Markov property $\mathbb{P}[X_k = x_k|X_{k-1} = x_{k-1}, \ldots, X_1 = x_1] = \mathbb{P}[X_k = x_k|X_{k-1} = x_{k-1}] = p(x_{k-1}, x_k)$ holds. Is it also true that $\mathbb{P}[X_k = x_k|X_{k+1} = x_{k+1}, \ldots, X_{k+N} = x_{k+N}] = \mathbb{P}[X_k = x_k|X_{k+1} = x_{k+1}]$ for $N > 1$? [2 pts]

b) SARSA learning algorithm is an *(...)*-policy algorithm? What does this mean? [1 pt]

c) Under what two conditions on the step-size α_k does the Q-learning algorithm converge almost surely to the true Q function? [1 pt]

d) In ϵ-greedy action selection, for the case of four actions and $\epsilon = \frac{1}{6}$, what is the probability that the greedy action is selected (assuming it is unique)? [1 pt]

e) What is the computational complexity of one iteration of value iteration (VI)? [1 pt]

f) In episodic RL problems, provide the definition of the sample complexity of an algorithm using the PAC framework. [1 pt]

g) Provide a minimax lower bound on the sample complexity in episodic RL problems. [1 pt]

h) From the answer of the previous question, motivate the use of ”function approximation” in RL. [1 pt]

i) In actor-critic algorithms, how many parameters do we need to update? What do they correspond to? [1 pt]

a) We have:

$$
\begin{aligned}
\mathbb{P}[X_k = x_k | X_{k+1} = x_{k+1}, \dots, X_{k+N} = x_{k+N}] &= \frac{\mathbb{P}[X_k = x_k, X_{k+1} = x_{k+1}, \dots, X_{k+N} = x_{k+N}]}{\mathbb{P}[X_{k+1} = x_{k+1}, \dots, X_{k+N} = x_{k+N}]} \\
&= \frac{\mathbb{P}[X_k = x_k] p(x_k, x_{k+1}) \dots p(x_{k+N-1}, x_{k+N})}{\mathbb{P}[X_{k+1} = x_{k+1}] p(x_{k+1}, x_{k+2}) \dots p(x_{k+N-1}, x_{k+N})} \\
&= \frac{\mathbb{P}[X_k = x_k] p(x_k, x_{k+1})}{\mathbb{P}[X_{k+1} = x_{k+1}]} \\
&= \frac{\mathbb{P}[X_k = x_k, X_{k+1} = x_{k+1}]}{\mathbb{P}[X_{k+1} = x_{k+1}]} \\
&= \mathbb{P}[X_k = x_k | X_{k+1} = x_{k+1}]
\end{aligned}
$$

b) SARSA is an on-policy learning algorithm. It means that the algorithm tries to estimate the value function of the policy run under the algorithm.

c) We need $\sum_k \alpha_k = \infty$ and $\sum_k \alpha_k^2 < \infty$ since Q-learning algorithm is a stochastic approximation algorithm.

d) Under ϵ-greedy policy, with probability $1 - \epsilon$ the greedy action is selected, and with probability ϵ a random action is selected. In particular the greedy action is selected with probability $1 - \epsilon + \epsilon/4 = 1 - 1/6 + 1/24 = 7/8$.

e) The VI algorithm requires in each iteration $\Theta(S^2 A)$ floating point operations.

f) The sample complexity SP of an algorithm is the number of episodes required so that the algorithm returns an ϵ-optimal policy with probability at least $1 - \delta$.

g) $\frac{T^2 SA}{\epsilon^2} \log(1/\delta)$ where T is the length of an episode.

h) The lower bound scales as SA, and hence for large state or action spaces, we need to work with function approximation.

i) We update a parameter for the (state, action) value function of the current policy (critic), and a parameter for the policy (actor).

Problem 2

A gambler has 2 SEK and needs to increase it to 10 SEK in a hurry. He can play a game with the following rules: a fair coin is tossed; if a player bets on the right side, he wins a sum equal to his stake, and his stake is returned; otherwise he loses his stake. The gambler decides to use a bold strategy in which he stakes all his money if he has 5 SEK or less, and otherwise stakes just enough to increase his capital, if he wins, to 10 SEK. Let $X_0 = 2$ and X_n be his capital after n throws.

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

b) What is its transition matrix? [2 pts]

c) Specify the communicating classes, and determine which ones are recurrent and which ones are transient. [2 pts]

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

(d) A supermarket has K cashiers. After shopping, you observe the queues for the different cashiers and have to decide which queue to join. Initially these queues have $n_1, \dots, n_K$ clients respectively (including the client being served). Time is slotted, and each client in service leaves with probability p at the end of the slot (her service is completed). We assume that you are the last client (no client joins or changes queues after you arrive). At the beginning of each slot, you observe the current states (lengths) of the various queues and may decide to switch queue (in which case you join the end of the new queue). You wish to maximize the probability to leave the supermarket before 100 slots. [5pts]

a) When the gambler has 2 SEK, he stakes 2 SEK, and gets 0 SEK w.p. 1/2 and 4 w.p. 1/2.
When he has 4 SEK, he stakes 4 SEK, and gets 0 SEK w.p. 1/2 and 8 w.p. 1/2.
When he has 8 SEK, he stakes x SEK to get 10 SEK if he wins. Hence $10 = 2x + (8 - x)$, and thus $x = 2$ SEK. If he loses, he gets 6 SEK w.p. 1/2 and 10 SEK w.p. 1/2.
When he has 6 SEK, he stakes x SEK to get 10 SEK if he wins. Hence $10 = 2x + (6 - x)$, and thus $x = 4$ SEK. If he loses, he gets 10 SEK w.p. 1/2 and 2 SEK w.p. 1/2.
The state space is hence $\{0, 2, 4, 6, 8, 10\}$.

b)

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

c) From a), the communicating classes are $\{0\}$, $\{2, 4, 6, 8\}$, $\{10\}$ (remember that two states are within the same class if we can go from one state to the other with positive probability). $\{2, 4, 6, 8\}$ is transient, and $\{0\}$, $\{10\}$ are recurrent.

d) We can model the problem as a finite-time horizon MDP with time horizon $T = 100$. The state $\emptyset$ represents the fact that you left the supermarket. At the beginning of a slot, if you haven't left the supermarket, the state should include the lengths of the various queues $x_1, \dots, x_K$ integers respectively less or equal to $n_1, \dots, n_K$. The action is the queue selected. i.e., $\mathcal{A} = \{1, \dots, K\}$. The rewards are always 0 before the end of the time horizon and 1 if you are in state $\emptyset$ at time 100. Hence for $t < 100$, $r_t(s, a) = 0$ for all state s and action a, and $r_T(\emptyset, a) = 1$ and $r_T(s, a) = 0$ for all a and $s \neq \emptyset$. The transition probabilities are as follows. Given a state $(x_1, \dots, x_K)$, we introduce $s^+(x_1, \dots, x_K) = \{k : x_k > 0\}$ (the set of non-empty queues). Now we distinguish two cases:

- If $a \in s^+(x_1, \dots, x_K)$ (you select to join the end of a non-empty queue). Then, if $\exists k : x'_k \notin \{x_k, x_k - 1\}$, we have:
$$p((x'_1, \dots, x'_K)|(x_1, \dots, x_K), a) = 0.$$
If $\forall k,\ x'_k \in \{x_k, x_k - 1\}$ (this is a possible transition), we have:
$$p((x'_1, \dots, x'_K)|(x_1, \dots, x_K), a) = p^A(1 - p)^B,$$
where $A = |\{k : x'_k = x_k - 1\}|$ and $B = |\{k : x'_k = x_k > 0\}|$. We also have:
$$p(\emptyset|(x_1, \dots, x_K), a) = 0.$$

- If $a \notin s^+(x_1, \dots, x_K)$ (you select an empty queue). Then again, if $\exists k : x'_k \notin \{x_k, x_k - 1\}$, we have:
$$p((x'_1, \dots, x'_K)|(x_1, \dots, x_K), a) = 0.$$
If $\forall k,\ x'_k \in \{x_k, x_k - 1\}$ (this is a possible transition), we have:
$$p((x'_1, \dots, x'_K)|(x_1, \dots, x_K), a) = p^A(1 - p)^B(1 - p),$$
where $A = |\{k : x'_k = x_k - 1\}|$ and $B = |\{k : x'_k = x_k > 0\}|$, and where the extra factor $(1 - p)$ means that although you are being served, you do not leave the supermarket. We finally have:
$$p(\emptyset|(x_1, \dots, x_K), a) = p.$$

- For all action a, $p(\emptyset|\emptyset, a) = 1$.

Problem 3

A race car must complete a total of T laps around a circuit. When equipped with new tires, the car can complete a lap in t_0 seconds and loses a random number of seconds $l(n)$ for each lap n in which it uses the same tires (i.e. to complete the m-th lap with the same tires it takes $t_0 + \sum_{i=1}^{m-1} l(i)$ seconds, with $l(i)$ drawn i.i.d. according to a Bernoulli distribution, $\mathbb{E}[l(i)] = 0.5$). At the start of each lap, the driver is allowed to execute a pit stop which takes 25 seconds (added to the lap's time) but allows the car to continue with fresh tires (i.e. it can complete a lap in t_0 seconds again). Design a pit stop strategy that will minimize the expected finishing time of the car.

a) Model this problem as a Markov Decision Process (describe this MDP in detail). [3pts]

b) Denote by $P(i)$, the lap when the last pit stop before lap i occurred and denote the wear of tires at lap i by $w_i = \sum_{j=P(i)}^{i-1} l(j)$. Establish that the optimal policy is threshold-based, i.e., at lap i, the optimal action is to make a pit stop if and only if $w_i \geq \alpha_i$ for some threshold α_i. [3pts]

c) Compute the optimal policy and expected completion time for the last 2 laps of the race for a wear level of w. [2pts]

d) Establish that under the optimal policy, the driver must never perform a pit stop in the last lap. [2pts]

a) The state at lap t can be described by t and the wear of the tires at the beginning of the coming lap, denoted by w. Hence $S = (\{1, ..., T\} \cup \{0, 1, ..., T\})$ (the first coordinate represents t, the second represents w).

The initial state is $(1, 0)$, and we denote by $V_t(w)$ the value of the MDP when starting the t-th lap with tires of wear w.

Actions are C or S (continue or stop).
When we decide to change the tires, the lap takes $t_0 + 25$ and if we do not change it takes $t_0 + w$. Hence the rewards are:

$$r((t, w), C) = -(t_0 + w), \quad r((t, w), S) = -(t_0 + 25).$$

Transition probabilities: $\mathbb{P}[(t + 1, w)|(t, w), C) = 1/2 = \mathbb{P}[(t + 1, w + 1)|(t, w), C)$, and $\mathbb{P}[(t + 1, 0)|(t, w), S) = 1$.

b) Bellman's equation relates V_i to V_{i+1}:

$$V_i(w) = \min(V_{i+1}(0) + t_0 + 25, 0.5(V_{i+1}(w) + V_{i+1}(w + 1)) + t_0 + w).$$

The first term in the 'min' corresponds to the action of executing a pit stop, whereas the second term corresponds to the decision to continue with the old tires. Hence, since $V_i(w)$ is increasing in w (can be shown by induction) we deduce that the optimal action at time i is to execute a pit stop if the wear w_i satisfies:

$$w_i \geq \alpha_i := \min\{w \in \{0, ..., T\} : 0.5(V_{i+1}(w) + V_{i+1}(w + 1)) + w \geq V_{i+1}(0) + 25\}. \tag{1}$$

The optimal strategy is hence threshold based.

c) For the last lap, we have

$$V_T(w) = \min(t_0 + 25, t_0 + w) = t_0 + \min(25, w)$$

.

Hence we do a pitstop if $\alpha_T = 25 \leq w$.

Next, we compute the value function at time $T - 1$.

$$V_{T-1}(w) = \min \left[V_T(0) + t_0 + 25, \frac{1}{2} \left(V_T(w) + V_T(w + 1) \right) + t_0 + w \right]$$

However, note that $V_T(0) = t_0$.

So,

$$V_{T-1}(w) = 2t_0 + \min \left(25, \frac{1}{2}(\min(25, w) + \min(25, w + 1)) + w \right)$$

Now we need to consider 2 different cases:
(i) if $w \geq 13 = \alpha_{T-1}$, we have $V_{T-1}(w) = 2t_0 + 25$ and it is optimal to change tires.
(ii) If $w \leq 12$, we have $V_{T-1}(w) = 2t_0 + 2w + \frac{1}{2}$, and it is optimal to continue.

d) In the penultimate lap, we do a pitstop if $w \geq 13$. This means that (if we follow the policy), the highest possible wear level we can enter the last lap with is 13. So, in the final lap $w \leq 13 < 25 = \alpha_T$. Hence we will never do a pitstop in the last lap.

Problem 4

Consider a discounted MDP with $\mathcal{S} = \{\text{A, B, C}\}$ and $\mathcal{A} = \{a, b, c\}$. We plan to use Q-learning or SARSA algorithm, and initialize the estimated Q-function as

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

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

$$(A, a, 5); (A, b, 10); (B, c, 7); (B, b, 4); (C, a, 2); (C, b, 1); (A, c, 10), \dots$$

where each triplet represents the state, the selected action, and the corresponding reward.

a) Provide the updated Q-values, if we are using the Q-learning algorithm, after these observations, assuming that the discount factor is $\lambda = 0.5$ and the learning rate is fixed to $\alpha = 0.1$. [3 pts]

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

c) Provide the updated Q-values, if we are using SARSA, after these observations, assuming that the discount factor is $\lambda = 0.5$ and the learning rate is fixed to $\alpha = 0.1$. [3 pts]

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

e) After these initial transitions, we run both algorithms under an ϵ-greedy policy, and use learning rates satisfying the almost sure convergence conditions of stochastic approximation algorithms. Q-learning and SARSA will converge to state-action value functions corresponding to different policies. Describe the characteristics of these. For SARSA, provide a set of equations that characterize the limiting state-action value function. [2 pts]

a)

$$Q^{(6)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 0.5 & 1 & 0 \\ 0 & 0.4 & 0.7 \\ 0.2 & 0.15 & 0 \end{bmatrix} \end{array} .$$

b) $\pi(A) = b$, $\pi(B) = c$, and $\pi(C) = a$.

c)

$$Q^{(6)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 0.5 & 1 & 0 \\ 0 & 0.4 & 0.7 \\ 0.2 & 0.1 & 0 \end{bmatrix} \end{array} .$$

d) $\pi(A) = b$, $\pi(B) = c$, and $\pi(C) = a$.

e) Q-learning converges to the true Q-function of the MDP, and the corresponding greedy policy is an optimal policy. SARSA converges to a (state, action) value function Q satisfying the fixed point equations: for all (s, ϕ) (ϕ denotes an action here)

$$Q(s,\phi) = r(s,\phi) + \lambda \sum_j p(j|s,\phi) \left[(1-\epsilon) \max_\psi Q(j,\psi) + \frac{\epsilon}{3} \sum_\psi Q(j,\psi) \right] .$$

If the above Q is used to output a policy, the latter would be the greedy policy w.r.t. Q.

Problem 5

All questions in this problem are concerned with episodic RL problems with episodes of length T steps. The state and the action at time t are denoted by s_t and a_t, respectively.

A *Bernoulli-logistic unit* is a stochastic neuron-like unit used in some Artificial NNs. Its input at time t is a feature vector $x(s_t)$; its output is a random action a_t having two values, 0 and 1, with $\Pr\{a_t = 1\} = P_t$ and $\Pr\{a_t = 0\} = 1 - P_t$ (the Bernoulli distribution). Let $h(s, 0, \theta)$ and $h(s, 1, \theta)$ be the preferences in state s for the unit's two actions given policy parameters θ. Assume that the difference between the action preferences is given by a weigthed sum of the unit's input vector, that is, assume that $h(s, 1, \theta) - h(s, 0, \theta) = \theta^T x(s)$, where θ is the unit's weight vector.

a) Assume the exponential soft-max distribution is used to convert action preferences to policies. Show that $P_t = \pi_\theta(s_t, 1) = 1/(1 + \exp(-\theta^T x(s_t)))$ (i.e., the logistic function). [2 pts]

b) Express the eligibility vector $\nabla \ln \pi_\theta(s, a)$ for a Bernoulli-logistic unit, in terms of a, and $x(s)$ [2 pts]

c) In which RL algorithm can we use $\sum_{t=1}^{T} \nabla \ln \pi_\theta(s_t, a_t)$ to update θ after observing an episode $(s_1, a_1, r_1, \ldots, s_T, a_T, r_T)$? [1 pt]

Now consider a general policy parametrization $\pi_\theta(s, a)$.

d) Prove that one can add a state-dependent bias term $b(s_t)$ to the policy gradient theorem without changing its expected value. [3 pts]

Hint: That is, prove that:

$$\mathbb{E}_{\pi_\theta} \left\{ \sum_{t=1}^{T} \nabla \log \pi_\theta(s_t, a_t) \big[\sum_{u=t}^{T} r(s_u, a_u) - b(s_t) \big] \right\} = \mathbb{E}_{\pi_\theta} \left\{ \sum_{t=1}^{T} \nabla \log \pi_\theta(s_t, a_t) \sum_{u=t}^{T} r(s_u, a_u) \right\}.$$

e) What could be a reason for adding such a bias term? How would you choose the bias term? [2 pts]

a) In the exponential soft-max policy, we have:

$$P_t = \frac{\exp(h(s,1,\theta))}{\exp(h(s,1,\theta)) + \exp(h(s,0,\theta))} = \frac{1}{1+\exp(h(s,0,\theta) - h(s,1,\theta))}$$

With $\theta^\top x(s_t) = h(s,1,\theta) - h(s,0,\theta)$, we get the desired result.

b) For $a = 1$:

$$\begin{aligned} \nabla \ln \pi_\theta(s,1) &= \nabla(-\ln(1+\exp(-\theta^\top x(s)))) \\ &= \frac{x(s)\exp(-\theta^\top x(s))}{1+\exp(-\theta^\top x(s))} = x(s)\pi_\theta(s,0) \end{aligned}$$

For $a = 0$:

$$\begin{aligned} \nabla \ln \pi_\theta(s,0) &= \nabla(-\ln(1+\exp(\theta^\top x(s)))) \\ &= -\frac{x(s)\exp(\theta^\top x(s))}{1+\exp(\theta^\top x(s))} = -x(s)\pi_\theta(s,1) \end{aligned}$$

Hence we have:

$$\nabla \ln \pi_\theta(s,a) = (-1)^{a-1} x(s)\pi_\theta(s,1-a).$$

c) REINFORCE

d) First note that the expected value can be expanded as

$$\begin{aligned} &\mathbb{E}_{\pi_\theta}\left\{\sum_{t=1}^{T} \nabla \log \pi_\theta(s_t,a_t)\big[\sum_{u=t}^{T} r(s_u,a_u) - b(s_t)\big]\right\} \\ &\quad = \mathbb{E}_{\pi_\theta}\left\{\sum_{t=1}^{T} \nabla \log \pi_\theta(s_t,a_t) \sum_{u=t}^{T} r(s_u,a_u)\right\} - \sum_{t=1}^{T} \mathbb{E}_{\pi_\theta}\Big\{\nabla \log \pi_\theta(s_t,a_t) b(s_t).\Big\} \end{aligned}$$

Hence, it is enough to show that

$$\mathbb{E}_{\pi_\theta}\big\{\nabla \log \pi_\theta(s_t,a_t) b(s_t)\big\} = 0.$$

We have that

$$\begin{aligned} \mathbb{E}_{\pi_\theta}\big\{\nabla \log \pi_\theta(s_t,a_t) b(s_t)\big\} &= \sum_{s,a} \Pr\{s_t = s, a_t = a\} \nabla \log \pi_\theta(s,a) b(s) \\ &= \sum_{s,a} \Pr\{a_t = a | s_t = s\} \Pr\{s_t = s\} \nabla \log \pi_\theta(s,a) b(s) \\ &= \sum_{s,a} \pi_\theta(s,a) \Pr\{s_t = s\} \nabla \log \pi_\theta(s,a) b(s) \\ &= \sum_{s} \Pr\{s_t = s\} b(s) \underbrace{\sum_{a} \pi_\theta(s,a) \nabla \log \pi_\theta(s,a)}_{=0} \\ &= 0, \end{aligned}$$

where the last factor is zero because

$$\begin{aligned} \sum_a \pi_\theta(s,a) \nabla \log \pi_\theta(s,a) &= \sum_a \pi_\theta(s,a) \frac{\nabla \pi_\theta(s,a)}{\pi_\theta(s,a)} \\ &= \sum_a \nabla \pi_\theta(s,a) \\ &= \nabla \sum_a \pi_\theta(s,a) \\ &= \nabla 1 \\ &= 0. \end{aligned}$$

e) Reduce the variance. Chose it as, for example, $\hat{v}_{\pi_\theta}(s)$.