STOCKHOLM UNIVERSITY
DEPT. OF MATHEMATICS
Div. of Mathematical statistics

MT 7051
EXAMINATION
8 Mar 2024

Exam in Reinforcement Learning
8 Mar 2024, time 08:00-13:00

Examinator: Chun-Biu Li, cbli@math.su.se.
Permitted aids: When writing the exam, you may use any literature. However, **Electronic devices are NOT allowed**

NOTE: The exam consists of 4 problems with 100 points in total. Logical explanation and steps leading to the final solution must be clearly shown in order to receive full marks.

NOTE: Your answers and explanations must be to the point, **redundant writing irrelevant to the solution will result in point deduction**.

NOTE: Mathematical notations in this exam are the same as those in the course book RLI

Problem 1 (Markov Decision Process and Dynamic Programming, total 25p)

a) Derive an expression for the state value $v_\pi(s)$ in terms of the action value $q_\pi(s,a)$ and the policy $\pi(a|s)$. **(3p)**

b) Show with explicit steps how to obtain the third line from the second line in Eq. 3.14 of the course book, i.e., prove that $E_\pi\left[R_{t+1}+\gamma G_{t+1}|S_t=s\right] = \sum_a \pi(a|s)\sum_{s',r} p(s',r|s,a)\big[r+\gamma E_\pi\left[G_{t+1}|S_{t+1}=s'\right]\big]$. **(8p)**

c) Derive an expression for the action value $q_\pi(s,a)$ in terms of state value $v_\pi(s)$, the model of the environment $p(s',r|s,a)$ and discount γ. **(8p)**

d) What is the difference between model-based and model-free methods? **(2p)** Explain clearly why the action value $q_\pi(s,a)$, instead of the state value $v_\pi(s)$, should be considered in model-free prediction and control. Please refer to the appropriate equation(s) in the course book to support your answer. **(4p)**

Problem 2 (Monte Carlo and Temporal Difference, total 24p)

a) Show that the ordinary average Eq. 5.5 equals to the weighted average Eq. 5.6 in the course book asymptotically, i.e., when the number of samples tends to infinity **(5p)**

b) Derive the update rules in Eq. 5.8 for the weighted average Eq. 5.7 in the course book. **(5p)**

c) In the Random-Walk example in p.125 of the course book, the reward is zero unless reaching the terminal state on the right. Explain why MC (whose value is updated only when the episode completes) and TD (whose value is updated in each step) result in different learning curves shown in the right figure in p.125. Justify your explanation with the appropriate equations. **(6p)**

d) Again in the Random-Walk example in p.125 of the course book, show with clear steps that the true state values are given by 1/6, 2/6, 3/6, 4/6 and 5/6 for the state A to E, respectively. **(8p)**

Problem 3 (n-step Bootstrapping and Function Approximation, total 25p)

a) The use of n-step bootstrapping (with $n > 1$) helps us look forward to future rewards and improve learning. As shown in Example 7.1 in the course book, the best performance of n-step bootstrapping usually occurs at intermediate values of n. Referring to Example 7.1, explain why the performance drops again when n becomes too large. Justify your explanation. **(7p)**

b) For function approximation of the action value $q(s, a, w)$ with w the function parameterization. Write down (no need to explain) the stochastic (or semi-) gradient descent update rules for w for (1) 1-step TD **(2p)**; (2) expected Sarsa **(2p)**; (3) n-step Sarsa **(2p)**; (4) λ-return **(3p)**; and (5) TD(λ) **(3p)**.

c) What equations are needed (beyond Eq. 10.10 in the course book) to specify the differential version of TD(0)? Justify your answer. **(6p)**

Problem 4 (Policy Gradient and DQN, total 26p)

a) Consider the policy parameterizations using the softmax function in Eq. 13.2 in the course book with linear action preference Eq. 13.3, show that the eligibility vector $\nabla \ln \pi(a|s, \theta) = \mathbf{x}(s, a) - \sum_b \pi(b|s, \theta)\mathbf{x}(s, b)$ **(7p)**

b) Consider the algorithm in the upper box in p.332 of the course book for the "One-step Actor-Critic TD(0) for episodic tasks". Can one generalize the method to One-step Actor-Critic Sarsa or Q-learning? If Yes, where in the box should be modified? If No, why? **(5p)**

c) Despite the success of DQN (Mnih et al. Nature 2015) in achieving superhuman performance in some of the ATARI games, propose TWO possible improvements (aside from the double DQN to remove the maximization bias) from what you learned from the course. Explain clearly with justifications **what the modifications to the update rules are** and **what aspects of the learning can be improved**. You can refer to the corresponding equations and sections in the course book in your justification. **(14p)**

Good Luck!