Indian Institute of Technology Kharagpur
Quiz 03 2021-22

Date of Examination: 29 Oct. 2021 **Duration:** 35 minutes
Subject No.: CS60077 **Subject:** Reinforcement Learning
Department/Center/School: Computer Science **Credits:** 3 **Full marks:** 20

Instructions

i. This question paper contains 3 pages and 2 questions. All questions are compulsory. Marks are indicated in parentheses. This question paper has been cross checked.

ii. Please write your name, roll number and date on top of the answer script.

iii. **Organize your work**, in a reasonably neat and coherent way. Work scattered all across the answer script without a clear ordering will receive very little marks.

iv. **Mysterious or unsupported answers will not receive full marks**. A correct answer, unsupported by calculations, explanation, will receive no marks; an incorrect answer supported by substantially correct calculations and explanations may receive partial marks.

v. In the online setting, you need to upload your answer scripts as **pdf file**. We will prefer a single pdf file. If you happen to have multiple files, please zip them and then uplaod as a single file. You can scan your worked out example or you can use latex to produce the pdf.

---

1. (a) (1 point) In incremental Monte-Carlo, we have seen the update rule for getting an online estimate of the state-value is,

$$V_T(S_1) = V_{T-1}(S_1) + \alpha_T \left(R_T(S_1) - V_{T-1}(S_1)\right) \tag{1}$$

where α_T is the learning rate at the T^{th} iteration. Write down the condition on α_T for the online estimate of V to converge to the true value in limit.

the estimate is going to converge to the true value, *i.e.*, $\lim_{T\to\infty}(S) = V(S)$, given two conditions that the learning rate sequence has to obey.

i) $\sum_T \alpha_T = \infty$

ii) $\sum_T \alpha_T^2 < \infty$

(b) (3 points) In tabular RL, we have seen 3 major types of back up rules - i) Dynamic Programing backup ii) Monte-Carlo backup and iii) TD backup. These 3 approaches differ in terms of their use of 'sample or full backups' and 'bootsrapping or no bootstrapping'. For the three approaches clearly write which one uses what type of backups (sample or full) and whether or not the use of bootstrapping is made.

Dynamic Programming: Full backups and bootstrapping.

Monte-Carlo: Sample backups and no bootstrapping

TD: Sample backups and bootstrapping

(c) (2 points) Write down the expression for the SARSA and expected SARSA update rules. Suppose, you are at the t^{th} step of the iteration and you want to update the value of $Q(s_t, a_t)$.

SARSA update rule is,

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left(R_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right) \tag{2}$$

Expwcted SARSA update rule is,

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left(R_{t+1} + \gamma \sum_{a \in \mathcal{A}} \pi(a/s_{t+1}) Q(s_{t+1}, a) - Q(s_t, a_t) \right) \tag{3}$$

(d) (1 point) Write down whether the following statement inside quotes is true or false. "SARSA is an on-policy algorithm."

True

(e) (1 point) Write down whether the following statement inside quotes is true or false. "Q-Learning is an on-policy algorithm."

False

(f) (1 point) Is bootstrapping used in the TD(λ) algorithm for $\lambda = 1$. Write 'yes' or 'no' as the answer.

no.

(g) (1 point) Which of the following is true regarding the usage of tabular value function estimation?

i. It does not generalize to new domains.
ii. It is hard to apply to continuous spaces.
iii. It requires a lot of memory.
iv. All of the previous.

iv. All of the previous.

(h) (1 point) The main problem in training DQNs using the standard stochastic gradient descent algorithm is due to

i. The dependence between the training instances *i.e.*, absence of i.i.d. assumption.
ii. High dimensional visual data i.e. use of frame images as inputs.
iii. Both the previous.
iv. None of the previous.

i. The dependence between the training instances *i.e.*, absence of i.i.d. assumption.

(i) (1 point) Write down whether the following statement inside quotes is true or false. "Although there may be multiple optimal policies, there is only one optimal value function for an MDP with bounded rewards, finite action and state spaces."

True

(j) (1 point) One of the issues in Q-Learning is the non-stationarity of the training examples which comes with bootstrapping. Which of the following takes care of such non-stationarity in part?

i. Experience Replay.
ii. Fixed Q-targets.
iii. Eligibity traces
iv. None of the previous.

ii. Fixed Q-targets.

(k) (1 point) Let $\mathcal{T} : \mathcal{V} \rightarrow \mathcal{V}$ be an operator. If $||\mathcal{T}u - \mathcal{T}v|| \leq L||u - v||$ with $0 \leq L < 1$ for any $u, v \in \mathcal{V}$, then $\mathcal{T}$ is,

i. a non-expansion.
ii. a contraction.

ii. a contraction.

(l) (1 point) Write down whether the following statement inside quotes is true or false. "If E_k denotes a $k-$step estimator, the E_1 is TD(1) and E_∞ is TD(0)."

False

2. (a) (2 points) Consider four different random variables X, Y, Z and W. If X and the joint random variables (Y, W) are conditionally independent given Z, then prove that X and Y are also conditionally independent given Z. *i.e.*, if $X \perp\!\!\!\perp (Y, W)|Z$ then prove that $X \perp\!\!\!\perp Y|Z$.

Hint: Try to show $P(X, Y|Z) = P(X|Z)P(Y|Z)$ where $P(X, Y|Z)$ can be written by marginalizing out W.

Solution: $X \perp\!\!\!\perp (Y, W)|Z$ means,

$$P(X, Y, W|Z) = P(X|Z)P(Y, W|Z) \tag{4}$$

$$\implies \sum_W P(X, Y, W|Z) = \sum_W P(X|Z)P(Y, W|Z) = P(X|Z)\sum_W P(Y, W|Z) \tag{5}$$

$$\implies P(X, Y|Z) = P(X|Z)P(Y|Z) \tag{6}$$

This means $X \perp\!\!\!\perp Y|Z$.

(b) (3 points) Consider again, four different random variables X, Y, Z and W. If X and the joint random variables (Y, W) are conditionally independent given Z, then prove that X and Y are also conditionally independent given the joint variable (Z, W). *i.e.*, if $X \perp\!\!\!\perp (Y, W)|Z$ then prove that $X \perp\!\!\!\perp Y|(Z, W)$.

Hint: Try to show $P(X, Y|(Z, W)) = P(X|(Z, W))P(Y|(Z, W))$. Start with writing down the definition of conditional probability of the expression in the left hand side of the previous line. In the numerator of the expression you wrote, use the conditional independence property of the problem statement. Then you have to use the result from part (b) above.

Solution:

$$\begin{aligned} P(X, Y|(Z, W)) &= \frac{P(X, Y, W|Z)}{P(W|Z)} \\ &= P(X|Z)\frac{P(Y, W|Z)}{P(W|Z)} \text{ [using relation (4) above]} \\ &= P(X|Z)P(Y|Z, W) \end{aligned} \tag{7}$$

We have seen from the previous problem that, $X \perp\!\!\!\perp (Y, W)|Z$ means $X \perp\!\!\!\perp Y|Z$ which in turn means $X \perp\!\!\!\perp W|Z$. So,

$$\begin{aligned} P(X|Z)P(W|Z) &= P(X, W|Z) \\ \implies P(X|Z) &= \frac{P(X, W|Z)}{P(W|Z)} = P(X|Z, W) \end{aligned} \tag{8}$$

Putting $P(X|Z)$ from eqn. (8) to eqn. (7), we get

$$P(X, Y|(Z, W)) = P(X|Z, W)P(Y|Z, W) \tag{9}$$