Exam: DAT440 / DIT471 Advanced topics in machine learning lp4 VT23

Examiner and teacher: Morteza Haghir Chehreghani, morteza.chehreghani@chalmers.se

- The final exam is in the form of a digital exam on 30 May 2023.
- The exam will take four hours.
- There are in total 34 questions: Most of the questions have one point/mark, and there are few questions with two or three points/marks. The mark of each question is specified.
- The total score is 40, which will be normalized to 60 and then the final grade will be computed as described in the course canvas page.
- The exam must be done individually, and you cannot use cheat-sheet or any other resources.
- You can have an ordinary calculator for the exam.
- The exam consists of only multiple-choice questions, where for each question, at least one option (and possibly more than one) can be correct. The correct number of options is specified for each question.
- There will be no negative score for wrong answers, but you need to choose all (and only) the correct answers to get the mark/score of the question.
- The examiner will visit the exam premises on two occasions to answer the clarification questions: i) one hour after the start of the exam, and ii) when an hour of the exam remains.
- You do not need to show your calculations for the questions.

1 Consider a stochastic multi-armed bandit problem with mean reward $\mu(a)$ for each arm $a \in \mathcal{A}$, where $\mathcal{A}$ is the set of arms, a^* is the optimal arm, a_t is the arm selected at time t, T is the time horizon, $\mathbb{P}$ is a prior distribution over problem instances and $\mathcal{J}$ is a problem instance. Choose the correct statements about the Bayesian regret $\mathrm{BR}(T)$ in this kind of problem.

[Number of correct options: 1]

- [ ] $\mathrm{BR}(T) = \mathbb{E}[R(T)|\mathcal{J}]$
- [ ] $\mathrm{BR}(T) = \mu(a^*) \cdot T - \mathbb{E}_{\mathcal{J}\sim\mathbb{P}}\left[\sum_{t\in[T]} \mu(a_t)\right]$
- [ ] $\mathrm{BR}(T) = \mathbb{E}_{\mathcal{J}\sim\mathbb{P}}\left[\mu(a^*) \cdot T - \sum_{a\in\mathcal{A}} \mu(a)\right]$
- [x] $\mathrm{BR}(T) = \mathbb{E}_{\mathcal{J}\sim\mathbb{P}}[\mathbb{E}[R(T)|\mathcal{J}]]$

Rätt. 1 av 1 poäng.

2 Consider a stochastic K-armed bandit problem with mean reward $\mu(a)$ for each arm $a \in \mathcal{A}$, where $\mathcal{A}$ is the set of arms and $|\mathcal{A}| = K$. We define, for an arbitrary arm $\hat{a} \in \mathcal{A}$, the observed average reward $\bar{\mu}_t(\hat{a})$ of arm $\hat{a}$ before round t. Moreover, $r_t(\hat{a})$ is the confidence radius at round t for arm $\hat{a}$, and $\mathrm{UCB}_t(\hat{a})$ and $\mathrm{LCB}_t(\hat{a})$ are the upper and lower confidence bounds, respectively, of arm $\hat{a}$ at round t. Choose the correct statements about the confidence bounds/radius.
[Number of correct options: 2]

- [ ] Under bad event, $\mathrm{LCB}_t(a) < \mathrm{LCB}_t(a')$ for $a, a' \in \mathcal{A}$, $a \neq a'$ always indicates $\mu(a) < \mu(a')$.
- [x] The confidence radius is a random variable because of randomness in the rewards and algorithm. ✓
- [x] Under the event $\mathcal{E} := \{\forall a \in \mathcal{A} \forall t \quad |\bar{\mu}_t(a) - \mu(a)| \leq r_t(a)\}$, $\mathrm{UCB}_t(a)$ provides a upper bound for $\mu(a)$. ✓
- [ ] If $\mathrm{UCB}_t(a) < \mathrm{UCB}_t(a')$ for arms $a, a' \in \mathcal{A}$ where $a \neq a'$, the mean reward of arm a' is larger than the mean reward of arm a.

Rätt. 1 av 1 poäng.

3 Consider the following game that proceeds over n rounds: In each round $t \in \{1, \ldots, n\}$, you choose either to play or do nothing. If you do nothing, then your reward is $X_t = 0$. If you play, then your reward is $X_t = 1$ with probability p and $X_t = -1$ otherwise.
In terms of regret, what is the optimal way of choosing actions, i.e., the best algorithm, when p is known?
[Number of correct options: 1]

- [ ] Always choose to do nothing.
- [ ] Always choose to play.
- [ ] If $p > 0.5$, choose to do nothing. Otherwise, choose to play.
- [x] If $p > 0.5$, choose to play. Otherwise, choose to do nothing. ✓

Rätt. 1 av 1 poäng.

4 Choose the correct statements about the regret in bandit problems.
[Number of correct options: 2]

- [ ] It is a random variables since the optimal arm in each round is not necessarily unique.
- [x] It is a random variable because of randomness in the rewards. ✓
- [x] It is a random variable because of randomness in the the algorithm. ✓
- [ ] It is a random variable because of randomness in the optimal arm in each round.

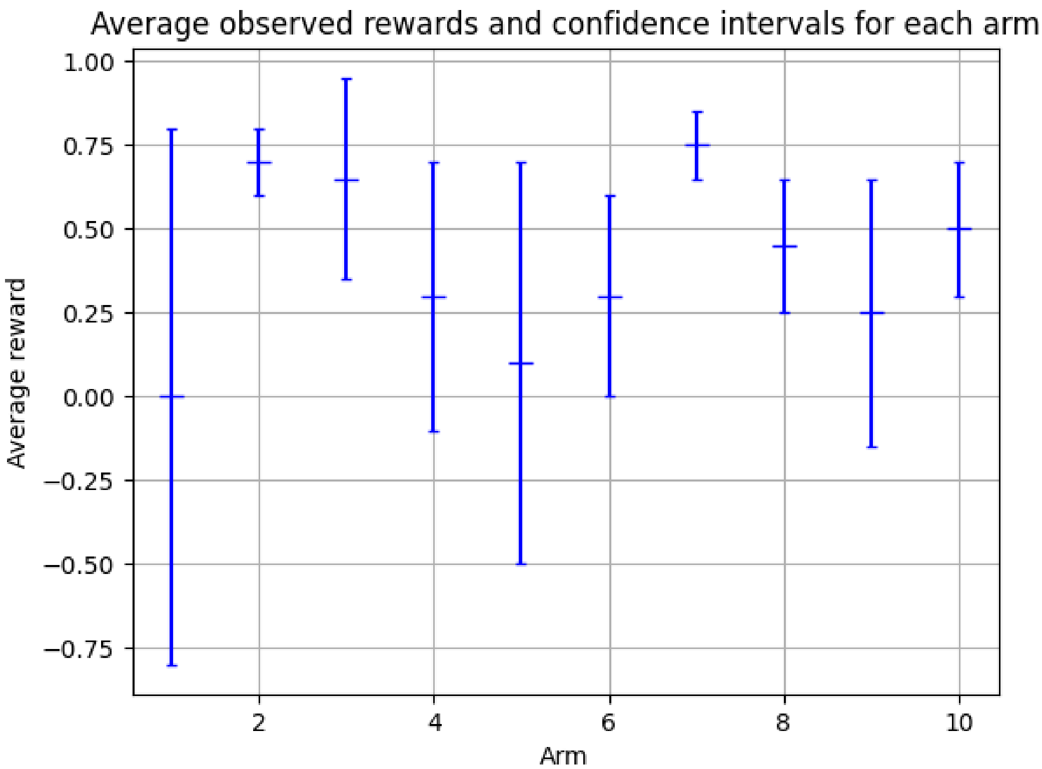
Rätt. 1 av 1 poäng.

5 Consider a stochastic K-armed bandit problem with mean reward $\mu(a)$ for each arm $a \in \mathcal{A}$, where $\mathcal{A}$ is the set of arms and $|\mathcal{A}| = K$. Choose the correct statements about the regret bound $\mathbb{E}[R(T)] \leq O(\log T)\left[\sum_{\text{arms } a \text{ with } \mu(a)<\mu(a^*)} \frac{1}{\mu(a^*)-\mu(a)}\right]$, where a^* is the optimal arm and T is the time horizon. [Number of correct options: 2]

- [x] It is instance-dependent.
- [x] Expectation is taken over the reward distribution and any randomness in the algorithm.
- [ ] It is instance-independent.
- [ ] It is a Bayesian regret.

Rätt. 1 av 1 poäng.

6



Consider a stochastic bandit problem with a set of arms $\mathcal{A}$, and $|\mathcal{A}| = 10$. For an arbitrary round, the figure displays the average reward $\bar{\mu}(a)$ up to this round for every arm $a \in \mathcal{A}$ and corresponding confidence intervals. Which arm(s) will algorithm UCB1 and Greedy choose in this round?

Algorithm UCB1 will choose [Arm 3] (Arm 9, Arm 5, Arm 4, Arm 10, Arm 1, Arm 7, Arm 3, Arm 6, Arm 8, Arm 2).

Greedy will choose [Arm 7] (Arm 7, Arm 5, Arm 4, Arm 10, Arm 6, Arm 2, Arm 9, Arm 1, Arm 3, Arm 8).

Rätt. 1 av 1 poäng.

7

1 Exploration phase: try each arm N times;
2 Select the arm $\hat{a}$ with the highest average reward (break ties arbitrarily);
3 Exploitation phase: play arm $\hat{a}$ in all remaining rounds.

Algorithm 1.1: Explore-First with parameter N.

Consider a stochastic K-armed bandit problem with a set of arms $\mathcal{A}$, and $|\mathcal{A}| = K$. Assume rewards are bounded in $[0, 1]$ and a total number of rounds T. Choose the correct statements about the Explore-First algorithm in bandit problems.
[Number of correct options: 2]

- [ ] If $K \gg T$, then the bad event can be always neglected.
- [ ] The regret is constant in the exploitation phase.
- [x] The cumulative regret in the exploration phase is upper bounded proportional to the number of arms K and the number of times N that each arm is tried in the exploration phase. ✔
- [x] Hoeffding's inequality can be used to define the clean event of the exploration phase. ✔

Rätt. 1 av 1 poäng.

8

Consider a Bayesian K-armed bandit problem with a set of arms $\mathcal{A}$, $|\mathcal{A}| = K$. Let $r_t \sim \mathcal{D}_{\mu(a_t)}$ denote the reward of arm $a_t \in \mathcal{A}$ chosen at round t, where the reward distribution $\mathcal{D}_{\mu(a_t)}$ is specified by its expected reward $\mu(a_t)$. Let $\mathbb{P}_H(\tilde{\mu}) := \Pr[\mu = \tilde{\mu} | H_t = H]$ denote the Bayesian posterior distribution after round t. Let H be a feasible t-history that is a concatenation of some feasible $(t-1)$-history H' and an action-reward pair (a, r), i.e., $H = H' \oplus (a, r)$. For a H-consistent algorithm, $\pi(a) = \Pr[a_t = a | H_{t-1} = H']$ is the probability that arm $a \in \mathcal{A}$ is chosen at round t given the history H'.

Assume that arm $a \in \mathcal{A}$ is chosen in round t and, subsequently, the reward r is observed. For a mean reward vector $\tilde{\mu} \in [0, 1]^K$, choose the correct statements about the joint distribution $\Pr[\mu = \tilde{\mu}, H_t = H]$.

[Number of correct options: 2]

- [x] $\Pr[\mu = \tilde{\mu}, H_t = H] = \Pr[H_{t-1} = H'] \times \Pr[\mu = \tilde{\mu}, (a_t, r_t) = (a, r) | H_{t-1} = H]$ ✔
- [ ] $\Pr[\mu = \tilde{\mu}, H_t = H] = \Pr[H_{t-1} = H'] \times \mathbb{P}_{H'}(\tilde{\mu}) \times \pi(a)$
- [ ] $\Pr[\mu = \tilde{\mu}, H_t = H] = \Pr[H_{t-1} = H'] \times \Pr[(a_t, r_t) = (a, r) | H_{t-1} = H]$
- [x] $\Pr[\mu = \tilde{\mu}, H_t = H] = \Pr[H_{t-1} = H'] \times \mathbb{P}_{H'}(\tilde{\mu}) \times \mathcal{D}_{\tilde{\mu}(a)}(r) \times \pi(a)$ ✔

Rätt. 2 av 2 poäng.

9 Table 1: Number of samples $n_t(a)$ and total observed reward (return) $s_t(a)$ of each a arm before round t.

a	1	2	2	3	5	6	7	8	9	10
$n_t(a)$	10	52	24	17	10	18	24	21	14	109
$s_t(a)$	0	36	10	5	0	6	10	8	3	92
$x_t(a)$	1	1	1	1	0	0	1	0	0	1

Consider a stochastic K-armed bandit problem with a set of arms $\mathcal{A}$, and $|\mathcal{A}| = K$. We define, at round t: $a_t \in \mathcal{A}$ is the arm played by the agent, $x_t(a) \in \{0, 1\}$ is the reward received by an agent if it plays arm $a \in \mathcal{A}$, $s_t(a) = \sum_{j=1}^{t-1} \mathbf{1}\{a_j = a\} x_j(a)$ is the cumulative reward of arm a before round t, and $n_t(a) = \sum_{j=1}^{t-1} \mathbf{1}\{a_j = a\}$ is the number of rounds that arm a has been played before round t. Consider the scenario in Table 1, where an agent has played in a multi-armed bandit environment with $K = 10$ arms up to round t. Note that the reward $x_t(a)$ is revealed to the agent *if and only if* arm $a \in \mathcal{A}$ is played in round t. With fixing the time horizon $T = 1000$, which arm will be chosen in rounds t and $t+1$ by the UCB1 algorithm with confidence radius $r_t(a) = \sqrt{\frac{2 \log T}{n_t(a)}}$?

Round t: Arm 6 (Arm 1, Arm 2, Arm 3, Arm 4, Arm 5, Arm 6, Arm 7, Arm 8, Arm 9, Arm 10)

Round $t+1$: Arm 2 (Arm 1, Arm 2, Arm 3, Arm 4, Arm 5, Arm 6, Arm 7, Arm 8, Arm 9, Arm 10)

Rätt. 2 av 2 poäng.

10 Consider a stochastic bandit problem with $K = 2$ arms with Gaussian rewards with means μ_1 and μ_2, respectively. Assume that the first arm is the optimal arm, i.e., $\mu^* = \mu_1$, and $\mu_1 = \mu_2 + \Delta$ with $\Delta > 0$. It can be shown that the regret of the Explore-First algorithm will be upper bounded by

$R(T) \leq \Delta\left(N + T\Phi\left(-\Delta\sqrt{\frac{N}{2}}\right)\right)$, where Φ denotes the cumulative distribution function (cdf) of the standard Gaussian distribution.

Provide a value N^* of N that minimizes the above regret upper bound.

Hint 1: $\frac{\partial}{\partial N}\Phi(a(N)) = \phi(a(N))\frac{\partial a(N)}{\partial N}$, where $\phi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$ denotes the probability density function (pdf) of the standard Gaussian distribution and $a(N)$ is an arbitrary function of N.
Hint 2: Use the Lambert function W, defined as for $y > 0$, $W(y)\exp(W(y)) = y$.
Note: $\lceil x \rceil$ is the **ceiling function** which maps x to the least integer greater than or equal to x.

[Number of correct options: 1]

- [ ] $N^* = \left\lceil \frac{2}{\Delta^2} W\left(\frac{T^2\Delta^4}{32\pi}\right) \right\rceil$ ✔
- [ ] $N^* = \left\lceil \frac{T^2\Delta^4}{32\pi} \right\rceil$
- [ ] $N^* = \left\lceil \frac{T^2\Delta^2}{16\pi} \right\rceil$
- [x] $N^* = \left\lceil \frac{1}{\Delta} W\left(\frac{T^2\Delta^2}{16\pi}\right) \right\rceil$ ✖
- [ ] $N^* = \left\lceil \frac{2}{\Delta^2} \right\rceil$
- [ ] $N^* = \left\lceil W\left(\frac{T^2\Delta^4}{32\pi}\right) \right\rceil$

Fel. 0 av 3 poäng.

11 Consider the Bellman optimality equations that yield v_*'s or q_*'s (i.e., the optimal state values or the optimal action values).
Now, given v_* or q_*, choose the correct statements.
[Number of correct options: 2]

- [x] One-step search or greedy search based on optimal state values is sufficient for an optimal policy in long-term.
- [ ] For an optimal policy in a state, one cannot rely only on the optimal state value at the respective state, and should consider the optimal state values at other states as well.
- [ ] For each state s, there will be always only one action at which the maximum is obtained in the Bellman optimality equation.
- [x] For each state s, there could be more than one action at which the maximum is obtained in the Bellman optimality equation.

Rätt. 1 av 1 poäng.

12 Consider the definition of return in reinforcement learning as the following.
$G_t \doteq \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$
Note that T indicates the horizon or the total number of time steps. Choose the correct statements.
[Number of correct options: 2]

- [ ] We have $G_t = G_{t+1} + \gamma R_{t+1}$
- [ ] We have $G_t = R_{t+1} + G_{t+1}$
- [x] We have either $T = \infty$ or $\gamma = 1$
- [ ] For $T = \infty$ we have $\gamma = 1$.
- [x] We have $G_t = R_{t+1} + \gamma G_{t+1}$

Rätt. 1 av 1 poäng.

13 Choose the correct statement(s) about the optimal policy in reinforcement learning.
[Number of correct options: 3]

- [x] Optimal policies share the same state-value function. ✔
- [ ] Optimal policies do not necessarily share the same state-value function.
- [x] Policy π is better than or equal to policy π' if and only if $v_\pi(s) \geq v_{\pi'}(s)$ for all states $s \in \mathcal{S}$. ✔
- [ ] Optimal policies do not always share the same optimal action-value function.
- [ ] Policy π is better than or equal to policy π' if and only if $v_\pi(s) \geq v_{\pi'}(s)$ for at least one of the states $s \in \mathcal{S}$.
- [x] Optimal policies share the same optimal action-value function. ✔

Rätt. 1 av 1 poäng.

14 What is the concept of "reward hypothesis" in reinforcement learning?
[Number of correct options: 2]

- [ ] It determines what should be achieved in a reinforcement learning task. ✔
- [x] It is equivalent to choosing actions with maximal expected/estimated reward at each time step. ✘
- [ ] It specifies how to reach the goals/purposes.
- [x] It indicates the goals and purposes can be modeled by the maximization of the expected value of the cumulative sum of rewards. ✔

Delvis rätt. 0 av 1 poäng.

15 Consider the policy evaluation (prediction) task in reinforcement learning to be performed under a MDP. The state values are shown by $v_\pi(s)$ under policy π for state s. Moreover, γ indicates the discount factor. Choose the correct statements about $v_\pi(s)$.
[Number of correct options: 3]

- [x] $v_\pi(s) = \mathbf{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s]$ ✔
- [x] $v_\pi(s) = \mathbf{E}_\pi[G_t|S_t = s]$ where G_t is the return for time t. ✔
- [ ] $v_\pi(s) = \mathbf{E}_\pi[R_{t+1} + \gamma v_\pi(S_t)|S_t = s]$
- [ ] $v_\pi(s) = \mathbf{E}_\pi[\gamma G_{t+1}|S_t = s]$ where G_{t+1} is the return for time $t+1$.
- [ ] $v_\pi(s) = \mathbf{E}_\pi[R_{t+1} + \gamma G_t|S_t = s]$ where G_t is the return for time t.
- [x] $v_\pi(s) = \mathbf{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s]$ where G_{t+1} is the return for time $t+1$. ✔

Rätt. 1 av 1 poäng.

16 Choose the correct statements about Dynamic Programming (DP) in reinforcement learning.
[Number of correct options: 3]

- [x] It may have great computation expense. ✔
- [ ] It is a model-free method.
- [x] It is based on a perfect model of the environment as a Markov decision process (MDP). ✔
- [ ] It learns from real experiences (interactions between the agent and environment).
- [ ] It always assumes the discount factor is 1.
- [x] It uses bootstrapping. ✔

Rätt. 1 av 1 poäng.

17 Choose the correct statements about Monte Carlo (MC) methods in reinforcement learning.
[Number of correct options: 2]

- [x] MC methods work based on averaging sample returns.
- [ ] MC methods need some prior knowledge of the actual environment's dynamics.
- [ ] MC methods perform the policy evaluation similar to Dynamic Programming methods.
- [x] MC methods learn from experience.
- [ ] Unlike Dynamic Programming, MC methods need an MDP.

Rätt. 1 av 1 poäng.

18 Consider the MC (Monte Carlo) methods in reinforcement learning. Which options propose a valid way to ensure all state-action pairs are visited?
[Number of correct options: 2]

- [ ] Using the discount factor of $\gamma < 1$
- [ ] Using greedy policy only with function approximation
- [x] Using an ϵ-greedy policy
- [x] Exploring starts
- [ ] Policy improvement by a greedy policy

Rätt. 1 av 1 poäng.

19 Consider the backup diagram for n-step TD (n-step Temporal Difference) Sarsa when n=2.
How many actions do appear in the backup diagram?
[Number of correct options: 1]

- [ ] 4
- [x] 3
- [ ] 2
- [ ] 5
- [ ] 1

Rätt. 1 av 1 poäng.

20 Choose the correct statement(s) about Temporal-Difference (TD) learning in reinforcement learning.
[Number of correct options: 3]

- [x] It uses bootstrap, similar to Dynamic Programming.
- [ ] It uses bootstrap, similar to Mote Carlo methods.
- [ ] It assumes the state transition probabilities are given.
- [ ] It assumes a model of the environment.
- [x] It is a model-free method.
- [x] It learns directly from raw experience (from interactions with the environment).

Rätt. 1 av 1 poäng.

21 Which item specifies the advantages of TD (Temporal Difference) over MC (Monte Carlo) in reinforcement learning?
[Number of correct options: 1]

- [ ] TD, unlike MC, does not require that the reward and next-state probability distributions are known.
- [x] TD, unlike MC, is naturally implemented in an online, fully incremental fashion.
- [ ] TD, unlike MC, learns from return samples.
- [ ] TD, unlike MC, does not require a model of the environment.

Rätt. 1 av 1 poäng.

22 Consider the following reinforcement learning algorithm.

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
  Initialize S
  Loop for each step of episode:
    Choose A from S using policy derived from Q (e.g., ε-greedy)
    Take action A, observe R, S'
    $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma \max_a Q(S', a) - Q(S, A)\big]$
    $S \leftarrow S'$
  until S is terminal

Choose the correct statement(s) about this algorithm.
[Number of correct options: 3]

- [x] This algorithm is a model-free method. ✔
- [ ] This algorithm is a model-based method.
- [x] This algorithm is for control (policy improvement). ✔
- [x] This algorithm is an off-policy method. ✔
- [ ] This algorithm does not improve the policy.
- [ ] This algorithm is an on-policy method.

Rätt. 1 av 1 poäng.

23 Consider n-step Expected Sarsa. Which option specifies its target update for state values?
R_{t+1} corresponds to reward at time $t+1$, γ is the discount factor, π is the policy, S_t is the state at time t, Q is the action value function, and a is an action.
[Number of correct options: 1]

- [ ] $R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n-1} + \gamma^n \sum_a \pi(a|S_t) Q_t(S_t, a)$.
- [ ] $R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n \sum_a \pi(a|S_{t+n}) Q_{t+n-1}(S_{t+n}, a)$. ✔
- [ ] $R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n \sum_a \pi(a|S_t) Q_{t+n-1}(S_t, a)$.
- [ ] $R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n \sum_a \pi(a|S_t) Q_t(S_t, a)$.
- [x] $R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n \sum_a \pi(a|S_{t+n}) Q_t(S_{t+n}, a)$. ✘

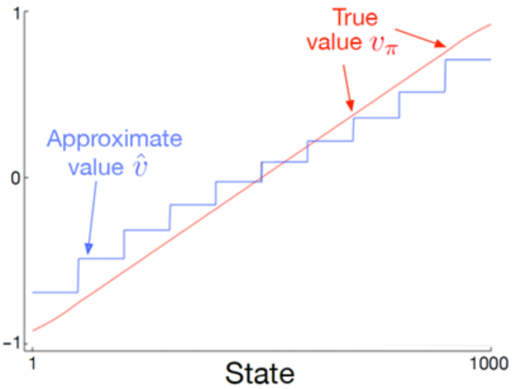
Fel. 0 av 1 poäng.

24 Consider function approximation for prediction in reinforcement learning, applied for state value estimation, i.e., $\hat{v}(S_t, \mathbf{w})$, where $\mathbf{w}$ corresponds to the parameters of the approximate function.
We use SGD (Stochastic Gradient Descent) to learn the parameters. We assume a linear function is used to approximate the state values, where $\mathbf{x}(S_t)$ specifies the features of state S_t. The update target is based on Monte Carlo (MC) shown by G_t.
Choose the correct statement about this problem.
[Number of correct options: 1]

- [ ] The SGD update is $\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha[G_t + \gamma\hat{v}(S_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, \mathbf{w}_t)]\mathbf{x}(S_t)$.
- [x] The SGD update is $\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha[G_t - \hat{v}(S_t, \mathbf{w}_t)]\mathbf{w}_t^{\mathrm{T}}\mathbf{x}(S_t)$.
- [ ] The SGD update is $\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha[G_t - \hat{v}(S_t, \mathbf{w}_t)]\mathbf{x}(S_t)$.
- [ ] The SGD update is $\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha[G_t + \gamma\hat{v}(S_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, \mathbf{w}_t)]\mathbf{w}_t^{\mathrm{T}}\mathbf{x}(S_t)$.

Fel. 0 av 1 poäng.

25 Consider function approximation for prediction in reinforcement learning, applied for state value estimation, i.e., $\hat{v}(S_t, \mathbf{w})$, where $\mathbf{w}$ corresponds to the parameters of the approximate function. We use the "state aggregation" method for $\hat{v}(S_t, \mathbf{w})$ along with SGD (Stochastic Gradient Descent) to learn the parameters.
We apply this method to a task with 1000 states (e.g., the 1000-state Random Walk problem). We obtain the following results about the estimated (approximate) and true state values.



Choose the correct statements about this problem.
[Number of correct options: 2]

- [ ] The update target used here is based on Monte Carlo (MC).
- [x] The update target used here is based on Temporal-Difference (TD).
- [ ] The update target used here is unbiased.
- [x] The update target used here is biased.

Rätt. 1 av 1 poäng.

26 Choose the correct statement(s) about function approximation in reinforcement learning.
[Number of correct options: 2]

- [x] It is useful for huge state spaces including visual images. ✔
- [ ] The approximate function must always be implemented using a deep neural network.
- [x] Compared to tabular reinforcement learning, it supports better the transfer of state values between similar states. ✔
- [ ] It is only applicable when the model of the environment is known.

Rätt. 1 av 1 poäng.

27 Which of the ideas is difficult to transfer easily from tabular reinforcement learning to deep reinforcement learning?
[Number of correct options: 1]

- [ ] Temporal Difference (TD)
- [ ] Double learning
- [ ] Experience replay
- [ ] Monte Carlo (MC)
- [x] UCB / Thompson sampling ✔

Rätt. 1 av 1 poäng.

28 Consider the DQN model designed for Atari games and choose the correct statements.
[Number of correct options: 2]

- [x] It uses experience replay. ✔
- [x] It clips the reward to be between -1 and +1. ✘
- [ ] It clips the TD error to be between -1 and +1. ✔
- [ ] It uses two completely independent neural networks instead of one.

Delvis rätt. 0 av 1 poäng.

29 Consider the episodic semi-gradient sarsa algorithm in reinforcement learning described as following.

Episodic Semi-gradient Sarsa for Estimating $\hat{q} \approx q_*$

Input: a differentiable action-value function parameterization $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Algorithm parameters: step size $\alpha > 0$, small $\varepsilon > 0$
Initialize value-function weights $\mathbf{w} \in \mathbb{R}^d$ arbitrarily (e.g., $\mathbf{w} = \mathbf{0}$)

Loop for each episode:
  $S, A \leftarrow$ initial state and action of episode (e.g., ε-greedy)
  Loop for each step of episode:
    Take action A, observe R, S'
    If S' is terminal:
      $\mathbf{w} \leftarrow \mathbf{w} + \alpha\big[R - \hat{q}(S, A, \mathbf{w})\big]\nabla\hat{q}(S, A, \mathbf{w})$
      Go to next episode
    Choose A' as a function of $\hat{q}(S', \cdot, \mathbf{w})$ (e.g., ε-greedy)
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha\big[R + \gamma\hat{q}(S', A', \mathbf{w}) - \hat{q}(S, A, \mathbf{w})\big]\nabla\hat{q}(S, A, \mathbf{w})$
    $S \leftarrow S'$
    $A \leftarrow A'$

How can the algorithm be converted to semi-gradient n-step sarsa?
[Number of correct options: 1]

- [ ] By replacing $R + \gamma\hat{q}(S', A', \mathbf{w})$ (if S' is not terminal state) by 0.
- [ ] By replacing R (if S' is not terminal state) by 0
- [x] By replacing R (if S' is not terminal state) by $G_{t:t+n}$, i.e., by n-step TD target. ✗
- [ ] By replacing $R + \gamma\hat{q}(S', A', \mathbf{w})$ (if S' is not terminal state) by $G_{t:t+n}$, i.e., by n-step TD target. ✓
- [ ] By replacing $\hat{q}(S', A', \mathbf{w})$ (if S' is not terminal state) by $G_{t:t+n}$, i.e., by n-step TD target.

Fel. 0 av 1 poäng.

30 Consider the one-step actor–critic method in reinforcement learning. Choose the correct statements about that.
[Number of correct options: 2]

- [x] The actor is responsible for learning the policy. ✓
- [ ] It uses the same update target as the REINFORCE algorithm.
- [ ] The critic is based on action value function.
- [ ] The policy improvement is based on ϵ-greedy.
- [x] The critic is based on state value function. ✓

Rätt. 1 av 1 poäng.

31 Consider the REINFORCE method with baseline, described below.

REINFORCE with Baseline (episodic), for estimating $\pi_{\theta} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Algorithm parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot,\boldsymbol{\theta})$
  Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
    $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ (G_t)
    $\delta \leftarrow G - \hat{v}(S_t,\mathbf{w})$
    Update $\mathbf{w}$
    Update $\boldsymbol{\theta}$

Choose the correct statement about this algorithm.
[Number of correct options: 1]

- [ ] The policy is based on a greedy or ϵ-greedy policy.
- [ ] The baseline must be constant and fixed for all episodes.
- [ ] The baseline term is δ.
- [x] The baseline term is $\hat{v}(S_t, \mathbf{w})$.

Rätt. 1 av 1 poäng.

32 Choose the correct statements about value-based and policy-based methods in reinforcement learning.
[Number of correct options: 2]

- [ ] ϵ-greedy is usually related to policy-based methods.
- [x] Actor-Critic methods learn both policy and value functions.
- [x] ϵ-greedy is usually related to value-based methods.
- [ ] ϵ-greedy is commonly used in both policy-based and value-based methods.
- [ ] ϵ-greedy is commonly used in Actor-Critic methods.

Rätt. 1 av 1 poäng.

33 Consider an episodic reinforcement learning problem where we apply Temporal Difference (TD). As you know, the TD error is defined as: $\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ where R_{t+1} is the reward at time $t+1$, γ is the discount factor, and $V(S_t)$ is the state value function for state S_t.

Assume that $V(.)$ does not change during the episode. Then, fill in the missing parts X and Y in the following expression (note that G_t is the return at time t).

$G_t = V(S_t) + \delta_t + \gamma X + \gamma^2(Y - V(S_{t+2}))$.

[Number correct options: 1]

- [ ] $X = G_{t+1}, Y = V(S_{t+1})$
- [ ] $X = G_{t+1}, Y = \delta_{t+1}$
- [ ] $X = \delta_t, Y = G_t$
- [ ] $X = V(S_{t+1}), Y = G_{t+2}$
- [x] $X = \delta_{t+1}, Y = G_{t+2}$ ✓

Rätt. 2 av 2 poäng.

34 Assume a model-free reinforcement learning problem based on Monte Carlo (MC) control wherein a refer to an action, s refers to a state, $\mathcal{A}(s)$ refers to the set of possible actions in state s, v is the state value function, and q is the action value function.

As you know, ϵ-greedy is used in Monte Carlo (MC) control. In particular, the ϵ-greedy policy π' is performed with respect to the actions values computed according to the old policy π (note that π is an ϵ-soft policy as well, i.e., with π, every action is selected with probability at least $\frac{\epsilon}{|\mathcal{A}|}$).

Choose the correct statements.

Note

[Number of correct options: 3]

- [ ] If $q_\pi(s, \pi'(s)) \geq \sum_a \frac{\pi(a|s) - \frac{\epsilon}{|\mathcal{A}(s)|}}{1-\epsilon} q_\pi(s,a)$ then we have $q_\pi(s, \pi'(s)) \geq v_{\pi'}(s)$
- [x] $q_\pi(s, \pi'(s)) = \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) + (1-\epsilon) \max_a q_\pi(s,a)$ ✓
- [x] If $q_\pi(s, \pi'(s)) \geq \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) + (1-\epsilon) \sum_a \frac{\pi(a|s) - \frac{\epsilon}{|\mathcal{A}(s)|}}{1-\epsilon} q_\pi(s,a)$ then we have $q_\pi(s, \pi'(s)) \geq v_\pi(s)$ ✓
- [ ] If $q_\pi(s, \pi'(s)) \geq \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s,a) + (1-\epsilon) \sum_a \frac{\pi(a|s) - \frac{\epsilon}{|\mathcal{A}(s)|}}{1-\epsilon} q_\pi(s,a)$ then we have $q_\pi(s, \pi'(s)) \geq v_{\pi'}(s)$
- [x] According to the policy improvement theorem, π' is an improvement over π or π' is as good as π. ✓
- [ ] $q_\pi(s, \pi'(s)) = \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_{\pi'}(s,a) + (1-\epsilon) \max_a q_{\pi'}(s,a)$

Rätt. 2 av 2 poäng.