Question 1. (5 points)

Agent receives rewards as follows:

$$r_k = \begin{cases} 0;\ \text{if } y_k = 0 \\ 1;\ \text{if } y_k = 1 \text{ and } k = 1 \\ 2r_{k-1} \text{with probability} \frac{1}{2};\ \text{if } y_k = 1 \text{ and } k > 1 \\ \frac{1}{2} r_{k-1} \text{with probability} \frac{1}{2};\ \text{if } y_k = 1 \text{ and } k > 1 \end{cases} \tag{1}$$

Determine $U^{y_{\leq 3}}$ of $y_1 = y_2 = y_3 = 1$.

Question 2. (10 points)

Agent receives rewards as follows:

$$r_k = \begin{cases} 0; \text{ if } y_k = 0 \\ 1; \text{ if } y_k = 1 \text{ and } k = 1 \\ r_{k-1} + 1 \text{ with probability} \frac{1}{2}; \text{ if } y_k = 1 \text{ and } k > 1 \\ r_{k-1} - 1 \text{ with probability} \frac{1}{2}; \text{ if } y_k = 1 \text{ and } k > 1 \end{cases} \tag{2}$$

Determine $U^{y_{\leq \infty}}$ of $y_{\leq \infty} = 1, 1, \ldots$ for $\gamma = \frac{1}{2}$.

Question 3. (5 points)

Consider classification with $Y = \{ 0, 1 \}$, where $P_c(y^*|x)$ is the probability that y^* is the true class of x, and rewards are given as

$$r_k = \begin{cases} 0 \text{ if } y = y^* \\ -1 \text{ if } y = 0 \text{ and } y^* = 1 \\ -3 \text{ if } y = 1 \text{ and } y^* = 0 \end{cases}$$

Consider the policy

$$y(x) = \arg\max_y P_c(y|x)$$

is this policy necessarily optimal, i.e. does it always coincide with the policy

$$\overline{\pi}(x) = \arg\max_y \mathbb{E}\left(r \mid x, y\right)$$

? Justify your answer mathematically.

Question 4. (2 points)

Discuss how the *exploration-exploitation* dilemma manifests itself in the *concept learning* scenario. Specify the conditions on which the execution of random actions would (would not) be useful for a concept-learning agent.

Question 5. (2 points)

Consider an algorithm that learns *monotone* disjunctions (or monotone conjunctions) from n-tuples of Boolean attribute values corresponding to n propositional variables. How can you use that algorithm to learn *general* disjunctions (or conjunctions) without changing it? You may change the number of inputs. How will your solution change the mistake bound in the case of the Winnow algorithm? Consider the number s of literals in the target disjunction constant.

Question 6. (1 points)

Let h, h' be propositional conjunctions. Is $h' \models h$ equivalent to $h \subseteq h'$? Justify your answer.

Question 7. (4 points)

Let h, h' be contingent propositional conjunctions that prescribe policies by

$$y = h(x) = 1 \text{ iff } x \models h$$

We say that h at least as general as h' if $h(x) = 1$ for any $x \in X$ such that $h'(x) = 1$. Is it true that $h' \models h$ if and only if h is at least as general as h'? Justify your answer.

Question 8. (15 points)

Consider an algorithm learning in the mistake bound model. Prove that if the condition $\sum_{k=1}^{\infty} |r_k| \leq poly(n_X)$ $(n_X \in \mathbb{N})$ is satisfied, then from some time $K \in \mathbb{N}$ the agent will not make any mistakes, i.e.,

$$\exists K \in \mathbb{N}, \forall k \in \mathbb{N} : k > K \rightarrow r_k = 0$$

Question 9. (3 points)

Give two examples of a non-contingent conjunction, one tautologically true and one tautologically false. Do the same for disjunctions. Explain how an incomplete truth assignment to n propositional variables is represented by a conjunction and decide whether such a conjunction may be tautologically false. Explain why Winnow does not learn from incomplete truth assignments.

Question 10. (5 points)

Show where the assumption that the target hypothesis is a *monotone disjunction* is needed in the proof of Winnow mistake bound and explain how the proof would fail if assuming a general target concept on $X = \{ 0, 1 \}^n$.

Question 11. (2 points)

Give the lgg for all pairs from

$$\mathcal{H} = \{\, \mathrm{p} \wedge \mathrm{q}, \mathrm{p} \wedge \neg \mathrm{q}, \mathrm{p} \vee \mathrm{q}, \mathrm{p} \vee \neg \mathrm{q} \,\}$$

whenever the lgg is defined for the pair.

Question 12. (3 points)

In concept classification, the generalization algorithm receives the sequence

$$x_1, x_2, \ldots, x_{10}$$

of observations (contingent conjunctions) where all x_k with odd indexes $(x_1, x_3, \ldots)$ are positive examples, and all the others are negative. The agent's sequence of hypotheses is

$$h_1, h_2, \ldots, h_{10}$$

so if some hypothesis is unchanged for m time steps, then there is a subsequence of m identical hypotheses above.

Determine

1. whether $h_2 = x_1$ or $h_2 = x_2$ or none of these options;
2. the sequence of hypotheses for the same agent that receives observations

$$x_1, x_1, x_3, x_3, \ldots, x_9, x_9$$

Question 13. (10 points)

Let h, h' be two propositional clauses or conjunctions. Show that $\mathsf{lgg}(h, h') = \mathsf{Lits}(h) \cap \mathsf{Lits}(h')$ is a least general generalization of h, h'.

Question 14. (15 points)

Determine if

1. $h \subseteq_\theta h'$
2. $h' \models h$

for

$$h = \mathrm{p}(x, y) \wedge \mathrm{p}(y, z) \wedge \neg \mathrm{p}(x, z)$$
$$h' = \mathrm{p}(\mathsf{a}, \mathsf{b}) \wedge \mathrm{p}(\mathsf{b}, \mathsf{c}) \wedge \mathrm{p}(\mathsf{c}, \mathsf{d}) \wedge \neg \mathrm{p}(\mathsf{a}, \mathsf{d})$$

Question 15. (5 points)

Consider the following statements

1. X = non-self-resolving FOL clauses
2. X = contingent FOL clauses
3. There is no $k \in \mathbb{N}$, $x \in X$ such that $h_k \models x$ and $h_k \not\subseteq_\theta x$, where h_k ($k \in \mathbb{N}$) are the hypotheses of the generalization algorithm.

Decide for each of the implications $1 \to 2$, $1 \to 3$, $2 \to 3$, whether it is true. Change the relation $h_k \models x$ in (3) so that all the implications you decided true are true when (1) and (2) assume conjunctions instead of clauses.

Question 16. (1 points)

Find two different least general generalizations of $\mathrm{p}(\mathsf{a})$ and $\mathrm{p}(\mathsf{b}) \vee \mathrm{p}(\mathsf{c})$, prove that they are indeed generalizations of the two clauses and prove that they are mutually θ-equivalent. Explain why two least general generalizations of the same pair of clauses or conjunctions must be θ-equivalent.

Question 17. (10 points)

Explain why the proof of the mistake bound n of the generalization algorithm is no longer valid when the assumption on X is changed to $X =$ non-self-resolving FOL conjunctions or non-self-resolving FOL conjunctions clauses, the relations $\subseteq, \subset$ are changed to $\subseteq_\theta, \subset_\theta$ (respectively), and we set $n = |\mathcal{P}|$. Show that the proof cannot be rectified, in particular that no finite mistake bound exists under said assumption even if $\mathcal{F} = \emptyset$.

Question 18. (3 points)

Determine the least general generalization of the following two assertions

1. *Superman is mortal or he is not a human.*
2. *Every human who smokes is mortal.*

by representing them as first-order logic clauses and computing their least general generalization with respect to the θ-subsumption order, and express the result in natural language.

Question 19. (2 points)

Let h, h' be FOL clauses and B a ground FOL conjunction. Show that if $h \subseteq_\theta h'$ then $h \subseteq_\theta^B h'$.

Question 20. (5 points)

Show that

$$h = \text{parent}(v_2, v_1) \wedge \text{male}(v_1) \rightarrow \text{son}(v_1, v_2)$$

and

$$g = \text{son}(v_1, v_2) \vee \neg\text{female}(\mathsf{a}) \vee \neg\text{parent}(\mathsf{a}, \mathsf{b}) \vee \neg\text{parent}(v_2, v_1) \vee \neg\text{male}(\mathsf{b}) \vee \neg\text{male}(v_1) \vee \neg\text{parent}(v_3, v_4) \vee \neg\text{parent}(\mathsf{b}, \mathsf{c}) \vee \neg\text{male}(v_4) \vee \neg\text{male}(\mathsf{c})$$

are equivalent relative to

$$B = \text{female}(\mathsf{a}) \wedge \text{parent}(\mathsf{a}, \mathsf{b}) \wedge \text{male}(\mathsf{b}) \wedge \text{parent}(\mathsf{b}, \mathsf{c}) \wedge \text{male}(\mathsf{c})$$

Question 21. (10 points)

Let

$$\begin{aligned} B =& \text{half}(4,2) \wedge \text{half}(2,1) \wedge \text{int}(2) \wedge \text{int}(1) \\ x_1 =& \text{even}(4) \\ x_2 =& \text{even}(2) \end{aligned}$$

1. Compute a least general generalization of x_1, x_2 observations relative to B .
2. Determine the reduction of the resulting clause relative to B and justify why it is indeed a reduction of it relative to B.

Question 22. (10 points)

Let X contain Herbrand interpretations for a finite set of $\mathcal{P}$ predicates and a finite set $\mathcal{F}$ of functions, and the observation complexity n_X be the tuple $(|\mathcal{P}|, |\mathcal{F}|)$. Show that the hypothesis class st-CNF (i.e., conjunctions of FOL clauses with at most s literals and at most t term occurrences in each literal) is learnable online from X.

Question 23. (2 points)

Consider a version-space agent whose initial hypothesis class $\mathcal{H}_1$ contains all non-contradictory conjunctions on 3 propositional variables.

1. Determine $|\mathcal{H}_1|$.
2. Give an upper bound on $|\mathcal{H}_2|$ given that $r_2 = -1$.

Question 24. (2 points)

Let $r_{\leq K}$ be a reward sequence of a standard agent and $h_{\leq K}$ be its sequence of hypotheses. Denote $M = \sum_{k=1}^{K} |r_k|$. Show that there is a hypothesis h retained for at least $\frac{K}{M+1}$ consecutive steps in $h_{\leq K}$, i.e.

$$h_{\leq K} = h_1, h_2, \ldots \underbrace{h, h, \ldots h}_{\text{at least } \frac{K}{M+1} \text{ times}} , \ldots h_K$$

Question 25. (5 points)

Let an agent PAC-learn $\mathcal{C}$ from X. Show that for any target concept from $\mathcal{C}$ on X, an arbitrary distribution $P(x)$ on X and arbitrary numbers $0 < \epsilon, \delta < 1$ and $K \in \mathbb{N}$, the condition $\mathsf{err}(h_K) \leq \epsilon$ with probability at least $1 - \delta$ implies that h_K is consistent with all observations in $x_{<K}$.

Question 26. (5 points)

Let X contain all real numbers from $[0; 1]$ which can be represented using 256 bits. Let $\mathcal{H} = X$, and the decision policy given by a $h \in \mathcal{H}$ is

$$h(x) = 1 \text{ iff } x > h$$

Determine a k such that with probability at least 0.9, $\mathsf{err}(h) < 0.1$, where h is an arbitrary hypothesis from $\mathcal{H}$ consistent with k i.i.d. examples from X. Estimate it using:

1. $\ln |\mathcal{H}|$
2. $\mathsf{VC}(\mathcal{H})$

Question 27. (3 points)

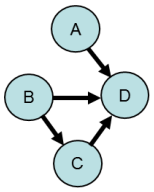
Give a real-life meaning to binary random variables A, B, C, for which the following Bayes graph is appropriate

1. $\mathbb{A} \to \mathbb{B} \to \mathbb{C}$
2. $\mathbb{A} \to \mathbb{B} \leftarrow \mathbb{C}$
3. $\mathbb{A} \leftarrow \mathbb{B} \to \mathbb{C}$

For each case, decide if the graph implies $A \perp\!\!\!\perp_P C \mid B$.

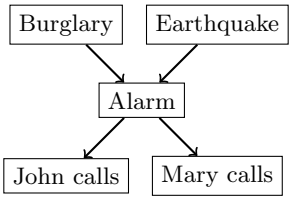
Question 28. (2 points)

How many parameters are needed in the Bayesian network below to fully specify a joint distribution on the random variables in vertices, which are binary?



Question 29. (5 points)

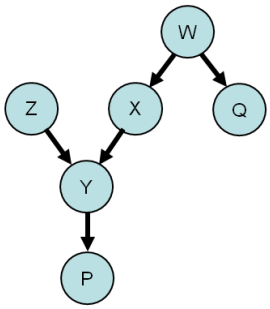
Given the Bayesian network below (conditional probability tables not shown),



1. (1 point) list all variables that 'John calls' is independent of, given that 'Alarm' is observed.
2. (4 points) express the probability that neither Mary nor John calls given that both burglary and earthquake happens, using only those (conditional) probabilities which are encoded in the conditional probability tables appropriate for this network. (Produce a formula referring to the random events by symbols such as A and their outcomes by symbols such as $b, \neg m$, etc.).

Question 30. (5 points)

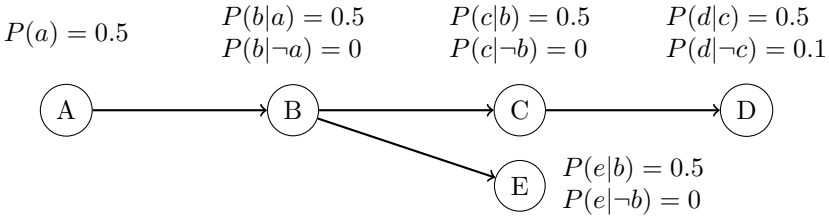
Let $X_1, X_2, \ldots \perp\!\!\!\perp_P Y_1, Y_2, \ldots \mid \mathcal{E}$ denote that $\forall i, j \in \{1, 2, \ldots\} : X_i \perp\!\!\!\perp_P Y_j \mid \mathcal{E}$. The empty set is denoted as $\emptyset$. Decide (true/false) for each of the statements below whether it is implied by the Bayes graph for P:



1. $Q \perp\!\!\!\perp_P X, Y, Z, P \mid W$,
2. $Z, Y, P \perp\!\!\!\perp_P W, Q \mid \emptyset$,
3. $Z \perp\!\!\!\perp_P X, W, Q \mid \emptyset$,
4. $Z \perp\!\!\!\perp_P X, W, Q \mid P$,
5. $Z, Y, P \perp\!\!\!\perp_P W, Q \mid X$,

Question 31. (15 points)

Consider the Bayes Network



1. Calculate $P(a|d)$ and $P(\neg a|d)$ by the factor method, i.e., using factor multiplication and variable elimination.
2. Determine $\arg\max_{A,B} P(A, B|d)$ by the factor method.

Question 32. (8 points)

Consider a Markov decision process.

1. (3 points) Explain why a fixed (independent of $x_1, x_2, \ldots$) sequence of actions as $y_1, y_2, \ldots$ ($y_k \in Y$) does not solve a Markov decision process, i.e. cannot guarantee optimality in reinforcement learning. In which field of AI would a fixed sequence of actions be an appropriate solution? Where is the boundary between game theory and reinforcement learning?

2. (2 points) Recall the value iteration algorithm. This algorithm is based on a general method for solving a set of equations. Give the name for this method and explain how it works in one or two sequences.

3. (3 points) Recall the policy iteration algorithm. This algorithm is based on another well known concept from machine learning and statistics. Name this algorithm and explain its idea in one or two sequences. Provide an example of other usages of this algorithm in computer science or mathematics.

Question 33. (5 points)

We state the Bellman equations the following way:

$$U(x) = r(x) + \gamma \max_{y \in Y(x)} \sum_{x'} P(x' \mid x, y) U(x').$$

In some literature, you may find under the same name a different equation:

$$U(x) = \max_{y \in Y(x)} \sum_{x'} P(x' \mid x, y)(r(x, y, x') + \gamma U(x'))$$

Describe in natural language the difference between the two formulations and decide if they are equally general, or else describe which one is the more general and why.

Question 34. (12 points)

Despite many reinforcement learning algorithms with additive rewards, it is common to use discounted rewards to model the environment.

1. (1 point) Give the range for the discount factor parameter.
2. (1 point) How does the agent behave when $\gamma = 0$?
3. (2 points) Give an example of an environment where a high discount factor is a good choice and why. Do the same for a low discount factor.
4. (3 points) In the case of the infinite horizon, discounting the rewards is necessary. Explain why.
5. (3 points) Imagine that your fancy reinforcement learning algorithm is not working on a chess game. Is it a good idea to include the discount factor in grid-search on meta-parameters of your algorithm? If yes, explain why; if not, would you still consider a range of discount parameter values and why?
6. (2 points) Suppose that
$$\max_{x \in X} |r(x)| = r_{\mathsf{max}}.$$
Using only r_{max} and γ, give the tightest lower and upper bound on the cumulative discounted reward in a single episode.

Question 35. (5 points)

Recall the learning rate parameter α of the temporal difference learning.

1. (1 point) Provide range for the α parameter.
2. (1 point) Explain the meaning of the α parameter.
3. (4 points) What must hold for α so that the temporal difference learning converges?
4. (1 point) Relate the temporal difference update rule

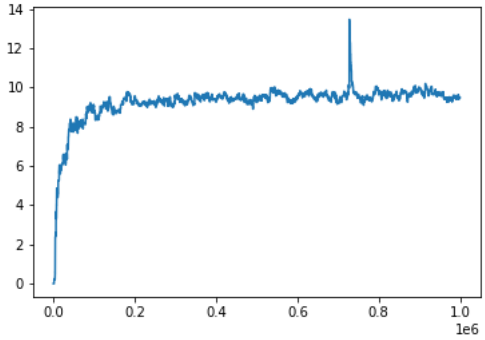
$$\widehat{U}(x) := \widehat{U}(x) + \alpha \left(r(x) + \gamma \cdot \widehat{U}(x') - \widehat{U}(x) \right)$$

to another well-known algorithm used in mathematical optimization.

Question 36. (13 points)

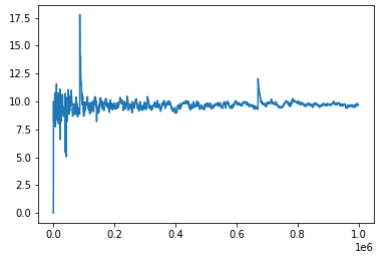
In this problem, we will study the influence of learning rate α on the value estimates $\widehat{U}$. All figures show learning of U using the temporal-difference method for the same state over one million episodes. The learning rate was selected so that the conditions for convergence were met.

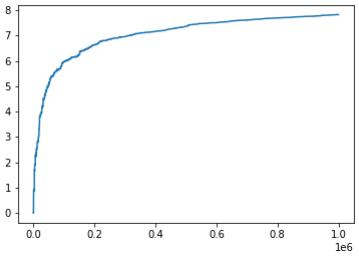
1. (2 points) Explain what causes the spike that you see around episode 700000.

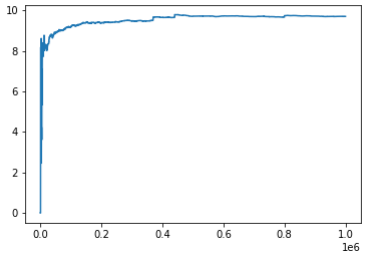


2. (2 points) Why do we need a different learning rate value for each state.

3. (6 points) Consider the following three scenarios of learning the value of a single state under a different learning rate. Explain which situation you consider optimal and identify when the learning rate was too small or too big. Propose a solution for the suboptimal cases.







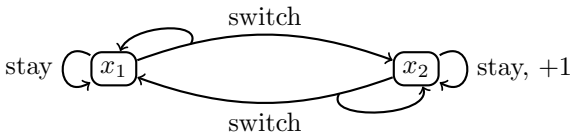
4. (3 points) The learning rate is a function of number of visits of a state $\alpha(n_x)$. Consider the following three functions

$$\alpha_1(n_x) = \frac{1}{10 + n_x}, \qquad \alpha_2(n_x) = \frac{3}{2 + n_x}, \qquad \alpha_3(n_x) = \frac{100}{99 + n_x}.$$

The figures in the question 3 were generated using those three learning rate functions. Match those functions to the figures and explain your decision.

Question 37. (10 points)

Consider the following MDP. Assume that reward is in the form $r(x, y)$, i.e., $r : X \times Y \mapsto \mathbb{R}$. Set $\gamma = \frac{1}{2}$.



Suppose that you have seen the following sequence of states, actions, and rewards:

$$x_1, \text{switch}, x_2, \text{stay}, +1, x_2, \text{stay}, +1, x_2, \text{switch}, x_1, \text{stay}, x_1, \text{switch}, x_1, \text{switch}, x_1, \text{stay}, x_1, \text{switch}, x_2, \text{stay}, +1, x_2$$

1. (4 points) What is $\widehat{U}^{\pi}(x_i)$ calculated by the Direct Utility Estimation algorithm?
2. (2 points) What is transition model P estimated by the Adaptive Dynamic Programming algorithm?
3. (2 points) In the ADP estimates, some of the rare events might have zero probability, even though they are possible. Provide a solution in which the rare events that the algorithm misses during learning have a non-zero probability.
4. (2 points) What are state values estimated by a Temporal Difference learning agent after two steps? Assume that $\alpha = 0.1$ and all values are initialized to zero.

[adapted from Richard Sutton's 609 course, see `http://www.incompleteideas.net/book/the-book-2nd.html`]

Question 38. (3 points)

Decide whether the following statement is true or false: *If a policy π is greedy with respect to its own value function U^π, then this policy is an optimal policy.* Explain your decision.

[adapted from Richard Sutton's 609 course, see `http://www.incompleteideas.net/book/the-book-2nd.html`]

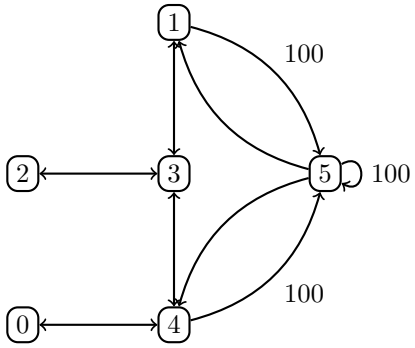
Question 39. (5 points)

Do the following exploration/exploitation schemes fulfill the 'infinite exploration' and 'greedy in limit' conditions? Which lead to the convergence of Q-values in Q-learning and which lead to the convergence of Q-values in SARSA. Does anything change if we are interested in the convergence of policy? $n_{x,y}$ denotes the number of times when action y was taken in state x. n_y is defined similarly.

1. a random policy
2. $$\pi(x) = \begin{cases} y, & \text{if } n_{x,y} \leq 100, \\ \arg\max_y Q(x,y), & \text{otherwise.} \end{cases}$$
3. ε-greedy policy with $\varepsilon = \frac{1}{n_x^2}$
4. ε-greedy policy with $\varepsilon = \frac{1{,}000}{999+n_x}$
5. ε-greedy policy with $\varepsilon = \frac{1}{\sqrt{n_x}}$

Question 40. (5 points)

Consider the following MDP with $\gamma = 0.8$, $r(5) = 100$, $r(\cdot) = 0$.



The initial matrix of Q-values is

$$\widehat{Q}(x,y) = \begin{bmatrix} - & - & - & - & 0 & - \\ - & - & - & 0 & - & 0 \\ - & - & - & 0 & - & - \\ - & 0 & 0 & - & 0 & - \\ 0 & - & - & 0 & - & 0 \\ - & 0 & - & - & 0 & 0 \end{bmatrix}.$$

Consider path $1 - 5 - 1 - 3$ and constant learning rate $\alpha = 0.1$. Show changes in Q values after the agent-environment interaction for the Q-learning algorithm.

[adapted from Richard Sutton's 609 course, see `http://www.incompleteideas.net/book/the-book-2nd.html`]

Question 41. (10 points)

Consider an active reinforcement learning algorithm implemented by SARSA or Q-learning.

1. (2 points) Unlike the temporal difference learning, SARSA and Q-learning algorithms learn Q values instead of U. Why is U not enough?
2. (3 points) Explain why those algorithms need to balance exploration vs. exploitation. What those terms mean, and which of those is preferred early in the learning.
3. (2 points) SARSA and Q-learning are guaranteed to converge to an optimal policy if both:
   - convergence criteria for learning rate α known from TD-learning are met, and
   - convergence criteria on the explore-exploit policy are met.

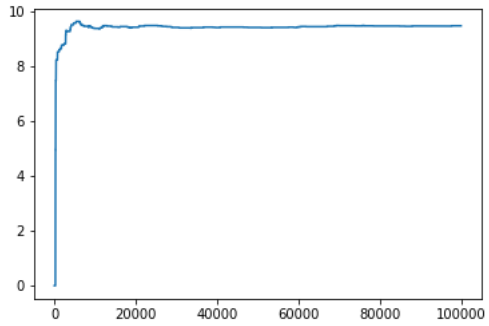
   What are those criteria placed on the explore-exploit policy?
4. (1 point) Provide an example of an explore-exploit policy that guarantees policy convergence for SARSA and Q-learning.
5. (2 points) Will one of the algorithms learn Q-values even if one of the conditions is not met? If yes, which and why, if not, explain.

Question 42. (10 points)

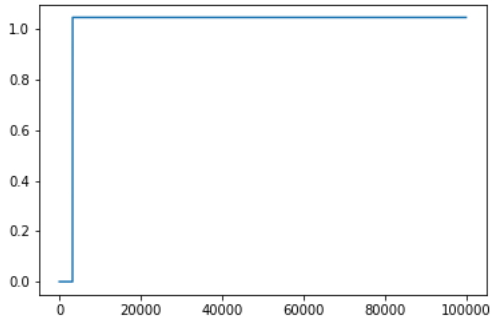
Consider an active reinforcement learning algorithm. You are not told whether it is an instance of SARSA or Q-learning. The implementation met all convergence criteria. All plots shown below are related to the same state-action pair Q-value, i.e., $\widehat{Q}(x, y)$. The action y is **suboptimal** in state x. The used explore-exploit policy was the ε-greedy policy, i.e., with probability ε a random action is selected; otherwise, the agent behaves greedily.

Now, consider four different situations of learning Q-values over 100 000 episodes.

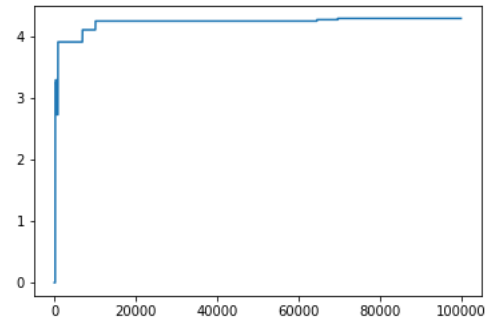
1.



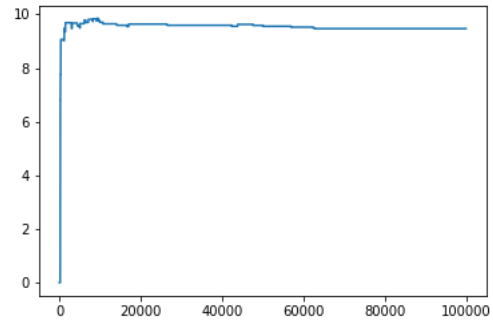
2.



3.

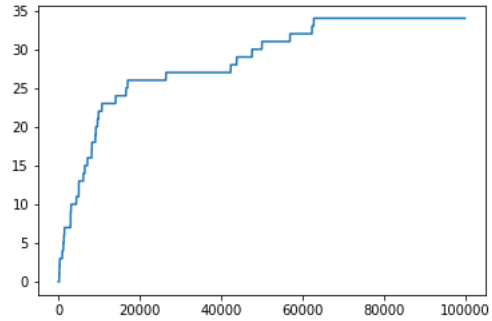


4.

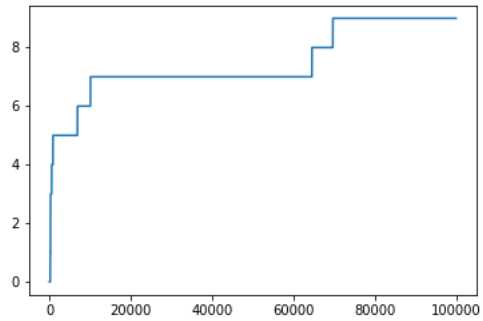


1. (4 points) Four plots below show how many times action y was selected by the agent in state x. For example, point $(1000, 6)$ means that the action y was selected 6 times in state x over the first 1000 episodes.

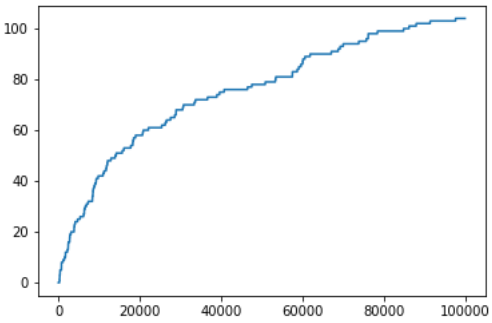
a.



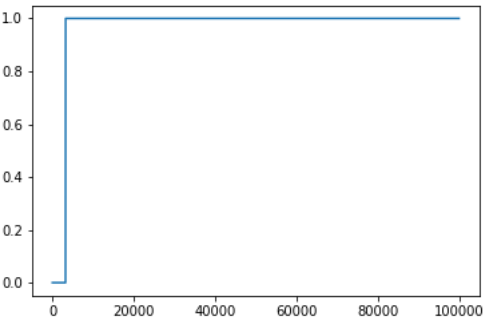
b.



c.



d.



Match figures a-d to figures 1-4. Explain your decision.

2. (2 points) The ε was set as a function of the number of visits of state x. Relate the following four functions to the figures 1-4.

   i. $\varepsilon(n_x) = \frac{8}{7+n_x}$ ii. $\varepsilon(n_x) = \frac{3}{2+n_x}$ iii. $\varepsilon(n_x) = \frac{100}{99+n_x}$ iv. $\varepsilon(n_x) = \frac{1000}{999+n_x}$

   Match those policies to figures 1-4. Explain your choice.

3. (1 point) Why should agents use different epsilon for different states.

4. (2 points) Decide whether the learning algorithm used was SARSA or Q-learning. Explain your decision.

5. (1 point) What is $Q(x, y)$? Explain your answer.