COMP0089PastExams

1 2022 Exam

Question 1

Consider a (stateless) bandit with m different actions $\{a1, ..., am\}$. When we select an action A_t at time t, a random reward R_t is drawn i.i.d. from a stationary distribution $p(R_t|A_t)$ that depends on the chosen action. These action-dependent reward distributions are unknown to the agent. As examples, consider the pseudo code for a random and greedy policy to interact with this bandit on a sequence of discrete steps.

Algorithm 1: Random Policy

while *True* **do**
  $A \leftarrow$ uniform from $\{a_1, \ldots, a_m\}$
  $R \sim p(\cdot \mid A)$
end

Algorithm 2: Greedy Policy (with constant step size $\alpha \in [0, 1]$)

$q_a \leftarrow 0\,, \forall a$
while *True* **do**
  $A \leftarrow \arg\max_a q_a$
  $R \sim p(\cdot \mid A)$
  $q_a \leftarrow q_a + \alpha(R - q_A)$
end

Figure 1

1.0.1 (a)

Question Using the format of the algorithms above, write down the pseudo code for 1) an ϵ-greedy policy and 2) a policy generated with the standard UCB algorithm (with which we mean the algorithm derived from a general Hoeffding concentration bound).

Answer:

Question Consider a bandit with two arms, a and b. So far, we have seen the following actions and rewards, on time steps $t = 1$ and $t = 2$:

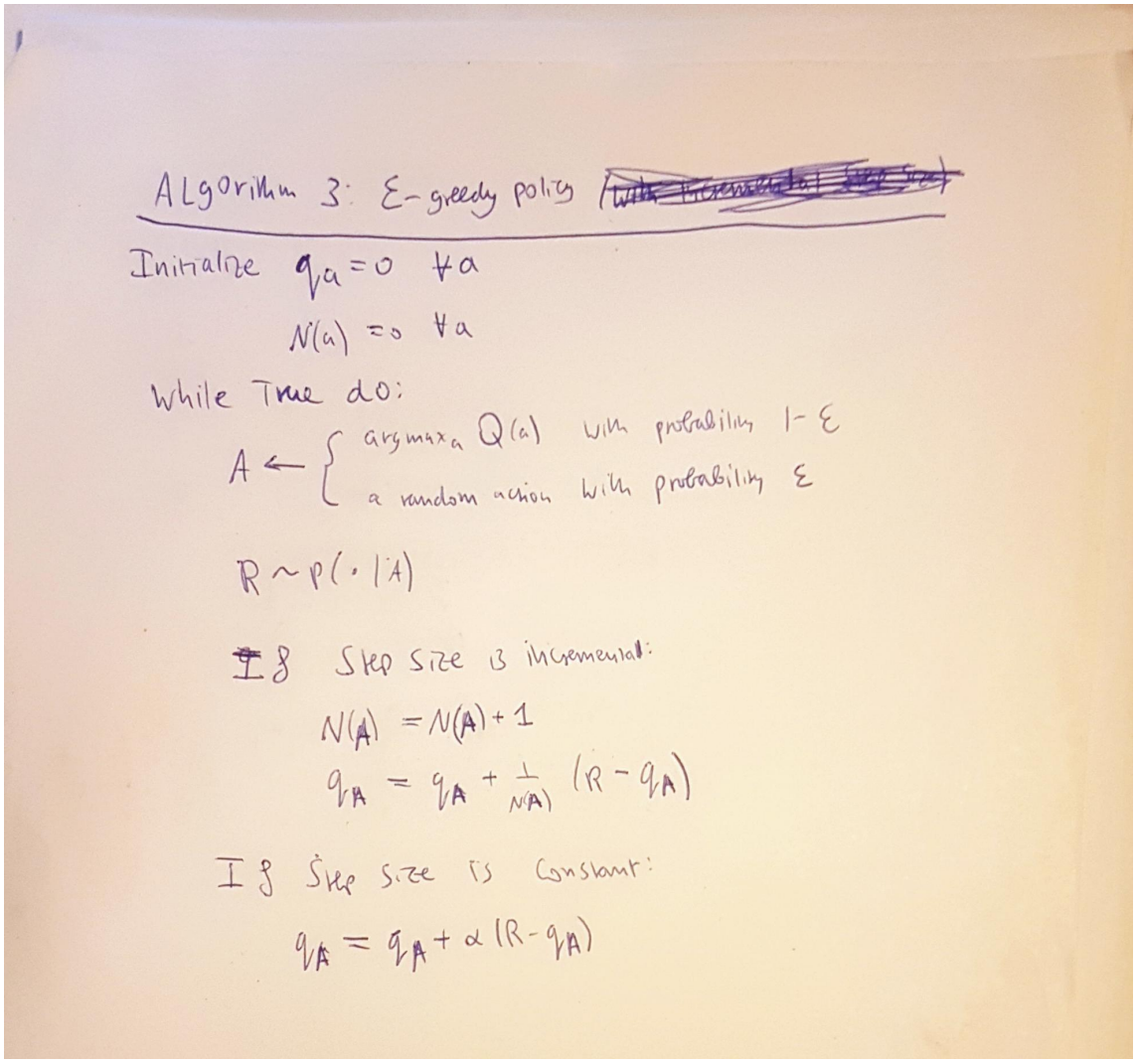
Algorithm 3: ε-greedy policy
Initialize $q_a = 0 \quad \forall a$
$N(a) = 0 \quad \forall a$
While True do:
$A \leftarrow$ { $\operatorname{argmax}_a Q(a)$ with probability $1-\varepsilon$; a random action with probability ε
$R \sim p(\cdot|A)$
If Step size is incremental:
$N(A) = N(A) + 1$
$q_A = q_A + \frac{1}{N(A)}(R - q_A)$
If Step size is Constant:
$q_A = q_A + \alpha(R - q_A)$

Figure 2

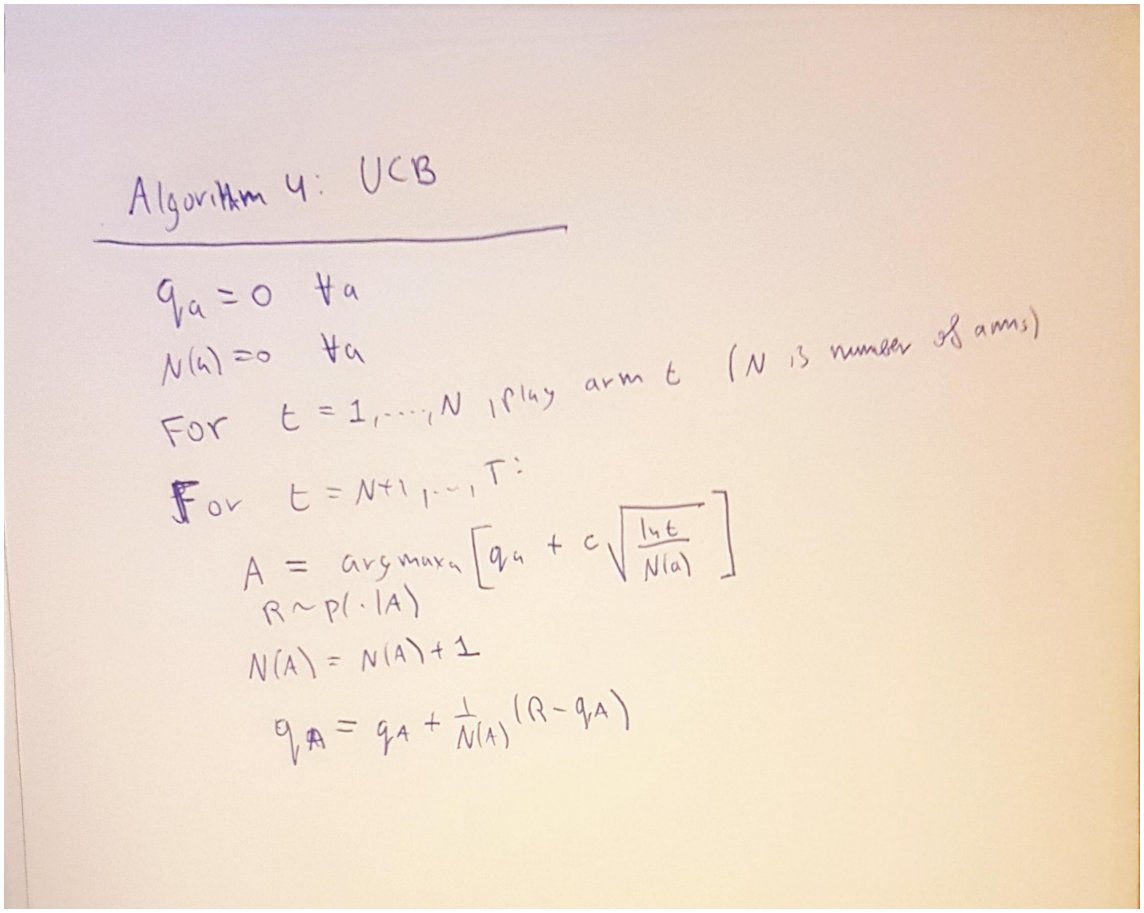
Algorithm 4: UCB
$q_a = 0 \quad \forall a$
$N(a) = 0 \quad \forall a$
For $t = 1, \ldots, N$ play arm t (N is number of arms)
For $t = N+1, \ldots, T$:
$A = \text{argmax}_a \left[q_a + c\sqrt{\frac{\ln t}{N(a)}} \right]$
$R \sim p(\cdot|A)$
$N(A) = N(A) + 1$
$q_A = q_A + \frac{1}{N(A)}(R - q_A)$

Figure 3

$$t = 1 : A_1 = a, R_1 = 0 \tag{1}$$
$$t = 2 : A_2 = b, R_2 = 1 \tag{2}$$

The rewards are knows to be Bernoulli random variables (so $R_t \in \{0, 1\}$) with unknown means.

Consider a Thompson sampling algorithm to select actions, with a uniform Beta prior at time $t = 0$ such that the probability density functions for the expected reward for both actions before seeing any data (at time $t = 0$) are defined by $p(E[R] = x|a) = 1$, for all $x \in [0, 1]$, and all a.

What is the probability under Thompson sampling of $A_3 = a$? Show your calculations, but keep it concise.

Answer:

We know that we have parameters α and β for the Beta distribution. For each action, we have (α_i, β_i) for $i \in \{1, 2\}$. When we get a reward of $+1$ for action 1, we update accordingly:

$$(\alpha_1, \beta_1) \leftarrow (\alpha_1 + R, \beta_1 + 1 - R) = ((\alpha_1 + 1, \beta_1 + 1 - 1) = (\alpha_1 + 1, \beta_1) \tag{3}$$

WE UPDATE ONLY THE ACTION THAT WAS TAKEN.

Now, in our case, the actions have the following posterior distributions (after the 2 steps):

$$\textbf{Step 1} : A_1 = 1, R_1 = 0 \implies (\alpha_a, \beta_a) \leftarrow (\alpha_a + R_1, \beta_a + 1 - R_1) \tag{4}$$
$$= (\alpha_a + 0, \beta_a + 1 - 0) = (1 + 0, 1 + 1 - 0) = (1, 2) \tag{5}$$
$$\textbf{Step 2} : A_2 = 2, R_2 = 1 \implies (\alpha_b, \beta_b) \leftarrow (\alpha_b + 1, \beta_b + 1 - 1) = (1 + 1, 1 + 1 - 1) = (2, 1) \tag{6}$$

Hence the actions had the following posteriors:

$$a \sim B(1, 2), b \sim B(2, 1)$$

Now, the density of actions a's posterior distribution is $B(1, 2)$. The density of a beta distribution when $\alpha, \beta \in \mathcal{N}$ is

$$\frac{\frac{x^{\alpha-1}(1-x)^{\beta-1}}{G(\alpha)\mathcal{G}(\beta)}}{\mathcal{G}(\alpha + \beta)} = \frac{\mathcal{G}(\alpha + \beta)x^{\alpha-1}(1 - x)^{\beta-1}}{\mathcal{G}(\alpha)\mathcal{G}(\beta)}$$

We have that $\mathcal{G}(\alpha) = (\alpha - 1)! \implies$ PDF of a is

$$\frac{x^{1-1}(1 - x)^{2-1}(2!)}{0!1!} = (1 - x)2 = 2 - 2x$$

and the PDF of b is

$$\frac{2!y^{2-1}(1 - y)^{1-1}}{0!1!} = 2y$$

Now, the question is, what is the probability that we choose a? i.e., what's the probability that $P(B(1, 2) > B(2, 1))$?

We have that the support of a Beta distribution is $x \in [0, 1]$. So we're searching for $P(B(1, 2) > B(2, 1)) = P(x > y)$:

$$p(x > y) = \int_0^1 \int_y^1 (2 - 2x)2y dx dy = \int_0^1 2y[2x - 2\frac{x^2}{2}]_y^1 dy \tag{7}$$
$$= \int_0^1 2y(2 - 1 - 2y + y^2)dy = \int_0^1 2y(1 - 2y + y^2)dy = \tag{8}$$
$$\int_0^1 2y - 4y^2 + 2y^3 dy = \frac{2y^2}{2} - \frac{4y^3}{3} + \frac{2y^4}{4} = 1 - \frac{4}{3} + \frac{1}{2} \approx 0.16 \tag{9}$$

(c)

Question Consider the UCB algorithm from question 1.a. Suppose we have two actions, a and b. Consider the initial exploration bonus for each to be infinite, as long as we have not selected the corresponding action, so that the algorithm first selects each action at least once. Suppose action a yields a Bernoulli random reward with $P(R = 1|a) = 1/3$ and $P(R = 0|a) = 2/3$. Action also yields a Bernoulli random reward, but with $P(R = 1|b) = 2/3$ and $P(R = 0|b) = 1/3$. What is the probability (before seeing any data) of selecting action a on the third time step (at which point we will have selected both a and b exactly one)? (Break ties uniformly, if relevant).

Answer:

So, the probability of selecting action a is the event where we select it with probability 1, given the reward for action a was bigger than that of b before time step 3. Or we select it randomly, given that before time step 3, the rewards of a and b were equal.

Mathematically: Define R_a and R_b to be the rewards that a and b yielded respectively before time step 3. So,

$$P(\text{selecting action a on time ste 3}) = \underbrace{P(\text{selecting a}|R_a > R_b)}_{=1} P(R_a > R_b) + \underbrace{P(\text{selecting a}|R_a = R_b)}_{=1/2} P(R_a = R_b)$$

Now, we have that $P(R_a > R_b) = P(R_a = 1)P(R_b = 0) = \frac{1}{3}\frac{1}{3} = \frac{1}{9}$
and $P(R_a = R_b) = P(R_a = 1)P(R_b = 1) + P(R_a = 0)P(R_b = 0) = \frac{1}{3}\frac{2}{3} + \frac{2}{3}\frac{1}{3} = \frac{4}{9}$
Hence,

$$P(\text{selecting action a on time ste 3}) = \frac{1}{9} + \frac{2}{9} = \frac{3}{9} = \frac{1}{3}$$

Question 2

1.1 (a)

(1) Question Write down the Bellman equation for state values $v_\pi(s)$ for a finite MDP, using explicit summations over states and actions (no expectation notation).

Answer:

$$v_\pi(s) = \sum_{a \in A(s)} \pi(a|s) \sum_{s',r} p(s', r|s, a)\big[r + \gamma v_\pi(s')\big] \tag{10}$$

(2) Question Write down the Bellman optimality equation for action values $q_*(s, a)$ for a MDP with continuous states and actions, using explicit integration (no expectation notation).

Answer: We have that the discrete case is:

$$\sum_{s',r} p(s', r|s, a)\big[r + \gamma_{a'} q_*(s', a')\big] \tag{11}$$

And the continuous case:

$$q_*(s, a) = \int_{s'} \int_r f_{s',r|s,a}(s', r|s, a)[r + \gamma \max_{a'} q_*(s', a')]drds' \tag{12}$$

(3) Question Write down the synchronous (across all states and actions) value iteration update rule for state values v_k, for a finite MDP (finite state and action space), using explicit summation (no expectation notation).

Answer:

$$v_{k+1}(s) = \max_a E[R_{t+1} + \gamma v_k(s_{t+1}|S_t = s, A_t = a] \tag{13}$$

$$= \max_a \sum_{s',a'} p(s', r|s, a)[r + \gamma v_k(s')] \tag{14}$$

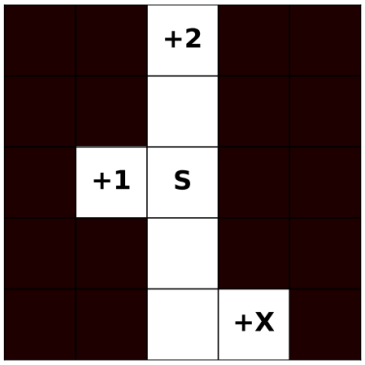

Figure 4

(b) Consider the MDP depicted in Figure 1. The agent starts at the cell marked by S.Whenever it enters a cell with a positive number, the agent receives that reward and the episode terminates. Consider $X = 3$.

(1) Question Draw the graph of the optimal value $v_*(s)$ of the starting state S, as a function of γ (on the x-axis, from 0 to 1) with $v_*(s)$ on the y-axis.

Answer:

Since we need to draw $v_*(S)$, this means that we need to find the values of the deterministic policies (as deterministic policies are the optimal). There can only be 3 policies: The one that goes to +1 with reward 1, the one that goes to +2 with reward $0 + \gamma 2 = \gamma 2$ and the policy that goes to +3 with reward $0 + \gamma 0 + \gamma^2 3 = \gamma^2 3$. We need to determine at which values of γ will the agent go to each state. e first compare when the agent is incentivized to go to state +3 rather than +2, which will be if the return of going to state +3 is greater than the return of going to +2. Mathematically:

$$\gamma^2(3) > \gamma(2) \tag{15}$$

$$\gamma > 2/3 \tag{16}$$

Now we compare when the agent will go to +2 instead of +1:

$$\gamma(2) > 1 \tag{17}$$

$$\gamma > 1/2 \tag{18}$$

So we have that for $\gamma \leq \frac{1}{2}$, agent chooses to go to +1. For $\frac{1}{2} < \gamma \leq \frac{2}{3}$, agent will go to +2, and for $\gamma \frac{2}{3}$, agent will go to +3.

And so the graph looks like this:

(2) Question When $X = 3$, what is the optimal policy in π_* state S as a function of the discount?

Answer:

As stated in the (1), the policy is: for $\gamma \leq \frac{1}{2}$, agent chooses to go to +1. For $\frac{1}{2} < \gamma \leq \frac{2}{3}$, agent will go to +2, and for $\gamma \frac{2}{3}$, agent will go to +3.

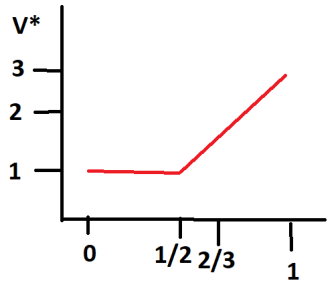

Figure 5

(c)

Suppose a behavioural scientist was doing an experiment where they gave rewards to an animal. Suppose the setting was a grid just like the MDP in Figure 4, with X = 5 (e.g., 5 food pellets at that location). It turns out that, after repeatedly exploring the grid, the animal seems to prefer going up to the reward of +2

(1) Question In this setting (with X = 5), prove that no scalar discount $[0, 1]$ exists for which the optimal policy is to go to +2.

Answer:

For the optimal policy to go to 2, we need the return for going to +2 to be greater than both the return of going to +1 and +5, i.e., mathematically

$$2\gamma > \gamma^2 5 \cap 2\gamma > 1 \tag{19}$$

$$\frac{2}{5} > \gamma \cap \gamma > \frac{1}{2} \tag{20}$$

Since $(\frac{1}{2}, \infty) \cap (-\infty, \frac{2}{5}) = \emptyset$, this means that there is no such γ for which the optimal policy is +2.
Consider a Monte Carlo return

$$G_t = R_{t+1} + f(R_{t+1} + f(R_{t+2} + f(...))).$$

(2) Question Standard discounting can be seen as applying a linear transformation $f(x) = \gamma x$, by multiplying the remaining return after each step by a factor γ. Consider the following alternative where instead of multiplying with a factor γ, we raise the value to power: $f(x) = x^\gamma$. Does this mathematical model better explain the observed behaviour, in the sense that a exists for which the optimal policy goes to +2? If so, give such a value for . If not, prove why not.

Answer: Again the return of +2 should be both greater than +1 and +5:

$$2^\gamma > (5^\gamma)^\gamma \cap 2^\gamma > 1 \tag{21}$$

$$\gamma \ln 2 > \gamma^2 \ln 5 \cap \gamma > 0 \tag{22}$$

$$\frac{\ln 2}{\ln 5} > \gamma \cap \gamma > 0 \tag{23}$$

$$0.43 > \gamma \cap \gamma > 0 \tag{24}$$

Hence for $0 < \gamma < 0.43$, the optimal policy will be to go to +2.

(d) Consider the true action values $q_{(s,a)}$, where these values are defined, as usual, as the expectation of the (standard) discounted cumulative rewards under a policy . Consider the optimal values to be defined, as usual, by $q_{(s,a)=\max_a q_\pi(s,a)}$.

(i) Question Is the greedy policy

$$\pi'(a|s) = \text{argmax}_a q_\pi(s,a) \tag{25}$$

always a policy improvement in the sense that $q_\pi(s,a) \geq q_\pi(s,a)$ for all s, a? Prove your answer.
Answer:
We have that

$$\begin{aligned} q_\pi(s,a) &= r(s,a) + \gamma E_{s'}[q_\pi(s', \pi(s'))] && (26)\\ &\leq r(s,a) + \gamma E_{s'}[\max_a q_\pi(s', a)] && (27)\\ &= r(s,a) + \gamma E_{s'}[q_\pi(s', \pi'(s'))] && (28) \end{aligned}$$

We also have that

$$q_\pi(s', \pi'(s')) \leq r(s', \pi'(s')) + \gamma E_{s''}[q_\pi(s'', \pi'(s'')] \tag{29}$$

Taking into consideration equations 16-19:

$$q_\pi(s,a) \leq E[r(s,a) + \gamma r(s', \pi'(s')) + \gamma^2 r(s^{''}, \pi'(s'')) + ...] \tag{30}$$

Where $r(s,a) + \gamma r(s', \pi'(s')) + \gamma^2 r(s'', \pi'(s'')) + ... = q_{\pi'}(s,a)$. Hence $q_\pi(s,a) \leq q_{\pi'}(s,a)$.
(ii) Question When does the equality $q_{\pi'}(s,a) = q_\pi(s,a)$ hold?
Answer:
$q_{\pi'}(s,a) = q_\pi(s,a)$ holds when $\pi = \pi'$

(e)

Question Consider the following dynamic-programming operator T:

$$\forall s, a : (Tq)(s,a) = r(s,a) + \gamma \max_b E_{S_{t+1} \sim p(|s,a)}[q(S_{t+1}, b)],$$

where $r(s,a) = E[R_{t+1}|S_t = s, A_t = a]$ is the expected reward (or deterministic reward) after taking action a in state s. When we apply this operator synchronously to all state-action pairs for some tabular initial action value function q, do the action values converge? If so, to which values does this procedure converge? Does the operator have a unique fixed point? Does it converge to the same fixed point as the standard Bellman optimality operator?

Answer:

This dynamic-programming operator is precisely the Bellman Optimality operator defined in the lecture notes. So it does converge and it converges to q_*. The fixed point corresponds to the action-value function q_* and it's the same as the Bellman optimality operator.