Exercises
Reinforcement Learning: Chapter 2

Exercise 2.1. In the comparison shown in Figure 2.1, which method will perform best in the long run in terms of cumulative reward and cumulative probability of selecting the best action? How much better will it be?

Exercise 2.3. Show that in the case of two actions, the softmax operation using the Gibbs distribution becomes the logistic, or sigmoid, function commonly used in artificial neural networks. What effect does the temperature parameter have on the function?

Exercise 2.4. Consider a class of simplified supervised learning tasks in which there is only one situation (input pattern) and two actions. One action, say $a$, is correct and the other, $b$, is incorrect. The instruction signal is noisy: it instructs the wrong action with probability $p$; that is, with probability $p$ it says that $b$ is correct. You can think of these tasks as binary bandit tasks if you treat agreeing with the (possibly wrong) instruction signal as success, and disagreeing with it as failure. Discuss the resulting class of binary bandit tasks. Is anything special about these tasks? How does the supervised algorithm perform on these tasks?

Exercise 2.5. Give pseudocode for a complete algorithm for the $n$-armed bandit problem. Use greedy action selection and incremental computation of action values with $\alpha = \frac{1}{k}$ step-size parameter. Assume a function $\text{bandit}(a)$ that takes an action and returns a reward. Use arrays and variables; do not subscript anything by the time index $t$. Indicate how the action values are initialized and updated after each reward.

Exercise 2.6. If the step-size parameters, $\alpha_k(a)$, are not constant, then the estimate $Q_k(a)$ is a weighted average of previously received rewards with a weighting different from that given by (2.7). What is the weighting on each prior reward for the general case?

Exercise 2.8. The results shown in Figure 2.4 should be quite reliable because they are averages over 2000 individual, randomly chosen 10-armed bandit tasks. Why, then, are there oscillations and spikes in the early part of the curve for the optimistic method? What might make this method perform particularly better or worse, on average, on particular early plays?

Exercise 2.9. The softmax action-selection rule given for reinforcement comparison methods (2.9) lacks the temperature parameter, $\tau$, used in the earlier softmax equation (2.2). Why do you think this was done? Has any important flexibility been lost here by omitting $\tau$?

Exercise 2.10. The reinforcement comparison methods described here have two step-size parameters, $\alpha$ and $\beta$. Could we, in general, reduce this to one parameter by choosing $\alpha = \beta$? What would be lost by doing this?
Exercise 2.12. An $\varepsilon$-greedy method always selects a random action on a fraction of the time steps. How about the pursuit algorithm? Will it eventually select the optimal action with probability approaching 1?

Exercise 2.13. For many of the problems we will encounter later in this book it is not feasible to update action probabilities directly. To use pursuit methods in these cases it is necessary to modify them to use action preferences that are not probabilities but that determine action probabilities according to a softmax relationship such as the Gibbs distribution (2.9). How can the pursuit algorithm described above be modified to be used in this way? Specify a complete algorithm, including the equations for action values, preferences, and probabilities at each play.

Exercise 2.15. The pursuit algorithm described above is suited only for stationary environments because the action probabilities converge, albeit slowly, to certainty. How could you combine the pursuit idea with the $\varepsilon$-greedy idea to obtain a method with performance close to that of the pursuit algorithm, but that always continues to explore to some small degree?

Exercise 2.16. Suppose you face a binary bandit task whose true action values change randomly from play to play. Specifically, suppose that for any play the true values of actions 1 and 2 are respectively 0.1 and 0.2 with probability 0.5 (case A), and 0.9 and 0.8 with probability 0.5 (case B). If you are not able to tell which case you face at any play, what is the best expectation of success you can achieve and how should you behave to achieve it? Now suppose that on each play you are told if you are facing case A or case B (although you still don’t know the true action values). This is an associative search task. What is the best expectation of success you can achieve in this task, and how should you behave to achieve it?
Exercise 2.17. Let \( r_1, r_2, \ldots, r_t \) be independent random variables, taking values in \([-R, R]\) and having a common expected value, \( Q^*\). Consider the incremental learning rule

\[
Q_{t+1} = (1 - \alpha_{t+1})Q_t + \alpha_{t+1}r_{t+1}, \quad Q_0 = 0,
\]

where the step-size, \( \alpha_t \), takes on values between 0 and 1: \( 0 \leq \alpha_t < 1 \). Let \( l_n = \sum_{k=1}^{n} \alpha_n \) and \( \beta_{kn} = \alpha_k \prod_{i=k+1}^{n} (1 - \alpha_i) \). (By convention, empty products are defined to take the value 1. That is, when \( k + 1 > n \), \( \beta_{kn} = \alpha_k \).) Show that for any \( \delta > 0 \),

\[
|Q_n - Q^*| \leq |Q^*| e^{-ln} + R \sqrt{2 \log \left( \frac{2}{\delta} \right) \sum_{k=1}^{n} \beta_{kn}^2}
\]

holds with probability at least \( 1 - \delta \). Argue that under the conditions \( \sum_{t=1}^{\infty} \alpha_t = \infty \) and \( \sum_{t=1}^{\infty} \alpha_t^2 < \infty \) the right hand side in the above expression converges to zero.

**Hint:** Start with the triangle inequality: \(|Q_n - Q^*| \leq |Q_n - \mathbb{E}[Q_n]| + |\mathbb{E}[Q_n] - Q^*|\). Show that \( Q_n = \sum_{k=1}^{n} \beta_{kn}r_k \) (empty sums take the value of 0). Next use Hoeffding’s inequality to compare \( Q_n \) and \( \mathbb{E}[Q_n] \) and then compare \( \mathbb{E}[Q_n] \) and \( Q^* \). Use the recursion \( A_{t+1} = (1 - \alpha_{t+1})A_t + \alpha_{t+1} \cdot 1 \), \( A_0 = 0 \), leading to \( A_n = \sum_{k=1}^{n} \beta_{kn} \) and bound the difference between \( \mathbb{E}[Q_n] \) and \( Q^* \). Along the way you will need to bound \( \prod_{k=1}^{n} (1 - \alpha_k) \). To do this consider \( \prod_{k=1}^{n} (1 - \alpha_k) = \exp(\sum_{k=1}^{n} \log(1 - \alpha_k)) \) and \( \log(1 + x) \leq x, \quad x > -1 \). For the second part, in order to show the convergence of \( \sum_{k=1}^{n} \beta_{kn}^2 \) to zero, use the Kronecker lemma, which follows from the Toeplitz lemma.

**Kronecker lemma:** Let \( x_k \in \mathbb{R} \), assume that \( \sum_k x_k \) converges. Let \( \sigma_k > 0, \sigma_k \to \infty, \sigma_{k+1} \geq \sigma_k \). Then

\[
\frac{1}{\sigma_n} \sum_{k=1}^{n} \sigma_k x_k \to 0, \quad n \to \infty.
\]

**Toeplitz lemma:** Let \( \{a_{ni}\}_{n,i \geq 1} \) be a double infinite array, \( \sup_n \sum_{i=1}^{n} |a_{ni}| \to +\infty \), \( \lim_{n \to \infty} \sum_{i=1}^{\infty} a_{ni} = a, \quad |a| < \infty \). Assume that for any \( i \), \( a_{ni} \to 0 \). Let \( b_n \) be a convergent sequence with limit \( b \). Then \( \sum_{i=1}^{n} a_{ni}b_i \to ab \) as \( n \to \infty \).

Note that the Toeplitz lemma can be proven by elementary steps and the Kronecker lemma follows from the Toeplitz lemma exploiting that since \( \sum_k x_k \) converges, the tail sequence \( v_n = \sum_{k=n+1}^{\infty} x_k \) is finite for all \( n \) and converges to zero as \( n \to \infty \). Hence \( x_k = v_{k-1} - v_k \) and then telescope. You should try to prove these yourself! They are extremely useful!

One typical use of the Kronecker lemma is to let \( \sigma_k = k \). In our case the logical choice is \( x_k = \alpha_k^2 \).

**Exercise 2.18.** Prove the following variant of the Toeplitz lemma: for \( a_n \geq 0 \) if \( \sum a_n \) diverges, \( x_n \to x \) then

\[
\frac{\sum_{i=1}^{n} a_i x_i}{\sum_{i=1}^{n} a_i} \to x, \quad n \to \infty.
\]
Exercise 2.19. Prove the following corollary to the second version of the Toeplitz lemma: Assume that $a_n \geq 0$ and that $\sum_i a_i$ diverges. Then

$$\frac{b_n}{a_n} \to c \Rightarrow \frac{\sum_{i=1}^n b_i}{\sum_{i=1}^n a_i} \to c$$

and

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^n b_i}{\sum_{i=1}^n a_i} = \infty \Rightarrow \limsup_{n\to\infty} \frac{b_n}{a_n} = \infty.$$

Hint:

$$\frac{\sum_{i=1}^n b_i}{\sum_{i=1}^n a_i} = \sum_{i=1}^n \frac{a_i}{\sum_{j=1}^n a_j} \frac{b_i}{a_i}.$$

Exercise 2.20. Consider the setting of Exercise 2.17. Assume that $\alpha_t = \alpha$, $0 \leq \alpha < 1$. Simplify the bound proven in Exercise 2.17 and then reason about the deviations of $Q_t$ from $Q^*$ for large $t$, when (i) $\alpha$ is “small” and when (ii) $\alpha$ is “big”.

Exercise 2.21. Consider again the setting of Exercise 2.17. Now assume that $\alpha_t = 1/t$. Simplify the bound proven in Exercise 2.17 and then reason about the deviations of $Q_t$ from $Q^*$ for large $t$. Now, think again: Hoeffding’s inequality could be used directly in this case since $Q_t$ is simply the sample average. How does the resulting bound differ from the bound that comes from the generic argument? How big is the difference? Where does the difference come from?

Exercise 2.22. For what values of $a$ does the sequence $\alpha_t = 1/t^a$ satisfy the Robbins-Monro (RM) conditions $\sum_{t=1}^\infty \alpha_t = \infty$, $\sum_{t=1}^\infty \alpha_t^2 < +\infty$?

Exercise 2.23. Show that convergence with probability one follows from the bound derived in Exercise 2.17 for the step-size sequence $\alpha_t = 1/t$, i.e., prove the law of large numbers under the condition that the random variables are independent and bounded. (Of course, the boundedness condition is not necessary for the law of large numbers to hold for a sequence of i.i.d. random variables. However, this is what we get with these simple tools.)

For a bonus prove the same under the conditions on the step-size sequence, $0 \leq \alpha_t \leq 1$, $\sum_{t=1}^\infty \alpha_t = \infty$, $\sum_{t=1}^\infty \alpha_t^2 < +\infty$,

$$C = \sup_n \frac{\alpha_{n+1}}{\alpha_n} < \infty, \rho = \sup_n \frac{\alpha_{n+1} - 1}{\alpha_{n+1}} < 2.$$

Argue that step-sizes of the form $\alpha_t = 1/t^a$ with $1/2 < a \leq 1$ satisfy this condition. (This result goes a little bit beyond the law of large numbers in that it concerns “weighted averages”.)
**Hint:** Prove that $Q_t$ converges *almost completely* to $Q^*$. This is sufficient since almost complete convergence implies convergence with probability one. You will need to prove that $\sum_n \exp(-c/z_n)$ is convergent, where $z_n = \sum_{k=1}^n \beta_{kn}^2$ and $c > 0$ is some constant. Remember that $\beta_{kn}$ took a special form when $\alpha_t = 1/t$. In this case $z_n$ has a special form that makes it easy to check that the above sum is convergent.

If you consider the problem for the bonus points, you want to look at the sequence $\{s_n\}_n$ defined by $\alpha_n s_n = z_n$. Exploiting that $z_{n+1} = (1 - \alpha_{n+1})^2 z_n + \alpha_{n+1}^2$, prove that $s_n$ is bounded and therefore due to $2x \log(x) \geq -1$ that holds for $x > 0$, $\sum_n c^{-2} \alpha_n^2 \geq \sum_n \exp(-c/\alpha_n)$, and so $\exp(-cz_n)$ is summable.

**Background:**

**Almost complete convergence:** We say that $\{X_n\}_n$ converges *almost completely* to $X$ if for any $\varepsilon > 0$, $\sum_n \mathbb{P}(\|X_n - X\| > \varepsilon) < +\infty$. In words, $X_n$ converges almost completely to $X$ if for any $\varepsilon > 0$, the tail probabilities of $X_n - X$ are summable.

**Almost complete convergence implies convergence with probability one:** Assume that $\{X_n\} \text{ converges almost completely to } X$. Then $X_n \to X$ with probability one. See e.g. the implication $4 \Rightarrow 2$ in the proof of (the missing) Proposition 2.8.1, pp. 59 in the lecture notes at http://www.math.harvard.edu/~knill/teaching/math144_1994/probability.pdf.

**Exercise 2.24.** Prove that for a random variable, $Z$, $\mathbb{P}(Z > \varepsilon) \leq Ce^{-c\varepsilon^2}$ for all $\varepsilon > 0$ ($C > 1, c > 0$) implies that

$$\mathbb{E}[Z] \leq \sqrt{1 + \log C \over c}.$$  

(This is Problem 8.2 in the book by Győrfi et al., “A Distribution free theory of nonparametric regression”, Springer, 2002.)

**Hint:** Use Jensen’s inequality:

**Jensen’s inequality** If $f$ is convex, then $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$. (Think of $f(x) = |x|$ to remember the direction of the inequality.)

Without the loss of generality we may assume that $Z \geq 0$ (why?). Then by Jensen’s inequality $\mathbb{E}[Z] \leq \sqrt{\mathbb{E}[Z^2]}$. Now use that for a non-negative random variable $U$, $\mathbb{E}[U] = \int \mathbb{P}(U > t) \, dt$ (why?), with $U = Z^2$. Split the resulting integral into two by breaking $[0, \infty)$ at $\log(C)/(c)$ and bound the integrand on the first part by 1, while on the second part use the assumption that we started with.

**Exercise 2.25.** Prove a bound on the expected error, $\mathbb{E}[|Q_n - Q^*|]$, under the RM conditions on the step-size and for large enough $n$.

**Hint:** Use the result of the previous exercise.
Exercise 2.26. Assume that the rewards in a stochastic bandit problem are bounded. Prove that if in $\varepsilon$-greedy, the amount of randomization is reduced to zero at an appropriate rate then $\Pr(a_t \neq a^*)$ converges to zero, where $a^* = \arg\max_a Q^*(a)$, assuming that there is a single such optimal action.

Hint: Use the second Borel-Cantelli lemma.

2nd Borel-Cantelli lemma: If $A_i$ is a sequence of independent events and $\sum_{i=1}^{\infty} \Pr(A_i) = \infty$ then the probability that an infinite number of these events hold is 1:

$$\Pr\left(\limsup_{n \to \infty} A_n\right) = 1.$$  
(Here, $\limsup_{n \to \infty} A_n \triangleq \bigcap_{n=1}^{\infty} \bigcup_{m \geq n} A_m$, so that if $\omega \in \limsup_{n \to \infty} A_n$, then it must be that for any $n$, $\omega \in A_m$ for some $m \geq n$, i.e., $\omega$ is in an infinite number of the events. It is also easy to see that all samples $\omega$ which are in an infinite number of the sets $A_n$ is also in the set $\limsup_{n \to \infty} A_n$.)

Exercise 2.27. Assume that we have a full information problem: In each time step we learn what rewards we could have achieved for all actions. Assume that the rewards are in the range $[-R, R]$ and the algorithm has been running for $t$ steps. Let $Q_t(a)$ be the average of rewards for action $a$. A sensible way of choosing actions is $a_t = \arg\max_a Q_t(a)$. Derive an upper bound on the probability that $\Pr(a_t \neq a^*)$ using Hoeffding’s inequality. What happens if there are multiple optimal actions?

Hint: Let $Q^*(a)$ be the expected payoff of action $a$, $Q^* = \max_a Q^*(a)$. Collect in a set $B$ the bad actions (this way you do not need to consider the two cases when there is a single optimal action or when there are multiple optimal actions): $B = \{a | Q^*(a) < Q^*\}$. Let $V$ be some real number. Argue that $\{\omega | a_t \in B\} \subset \{\omega | \forall a^* \notin B, Q_t(a^*) \leq V \text{ or } \exists a \in B \text{ s.t. } Q_t(a) > V\}$. Use that from $A \subset B$ it follows that $\Pr(A) \leq \Pr(B)$ and that $\Pr(B_1 \cup \ldots \cup B_k) \leq \sum_{i=1}^{k} \Pr(B_i)$. Consider choosing $V$ in the interval $(\max_{a \in B} Q^*(a), Q^*)$ and then use Hoeffding’s inequality.
Exercise 2.28. (Programming exercise) Experiment with 2-armed Bernoulli bandits! In the case of a Bernoulli bandit the payoff distribution of arm $a$ is concentrated on $0, 1$: and $\mathbb{P}(r_t = 1|a_t = a) = p(a)$, the probability that we see a reward of one is the same as the expected payoff. Now consider the case when there are only two arms, let’s call them 1 and 2 (so $a \in \{1, 2\}$). Such Bernoulli bandits are perfectly determined by two parameters, $p, q$: $p(1) = p, p(2) = q$. Implement $\varepsilon$-greedy with constant and decreasing $\varepsilon_t$, with $\varepsilon_t = \min(c/t, 1)$, $c > 0$ with various values of $c > 0$. Implement UCB (choose $p > 2$, where $p$ now denotes the parameter of UCB). Explore this space of bandit problems by plotting an estimate of the expected regret after $n = 10^2, 10^4, 10^6, 10^8$ time steps for selected choices of the parameters of the algorithm(s). You can estimate the expected regret by running multiple times the same algorithm. Pay attention to the variance of your estimates. Which algorithm would you recommend and how does the parameter settings influence the results? The recommendation should discuss how $(p, q)$ and the choice of the algorithm/parameters interact. What algorithm would you recommend if the parameters of the bandits were drawn uniformly at random from the unit square? Summarize your conclusions in a paragraph of few sentences.

**Hint:** Remember that we are interested in the expected regret. One good way of reducing the variance of the estimate of the expected regret is to use with the expected payoffs of the arms when computing the estimates. As an experimenter you have access to the expected payoffs of the arms, but the bandit algorithm will not.