EN.520.637: Foundations of Reinforcement Learning Homework 5

Ha Manh Bui (CS Department)
hbui13@jhu.edu

Fall 2023

1 Problem 1

For the multi-arms bandit problem we discussed in the class. Suppose that we get return G_n at n-th time we do action a, and $\mathbb{E}[G_n] = r$, $n = 1, 2 \cdots$.
Let Q_{n+1} be our estimates of r after we do action a the n-th time, and we have the following update rule

$$Q_{n+1} = Q_n + \alpha_n(G_n - Q_n), \quad Q_1 = 0.$$

We define $V_n = \mathbb{E}\left[(Q_n - r)^2\right]$.
(a) (Decreasing step size) Let $\alpha_n = \frac{1}{n}$, show that
(i) $Q_{n+1} = \frac{1}{n}\sum_{i=1}^{n} G_i$, $n = 1, 2, \cdots$.

Proof. Since $Q_{n+1} = Q_n + \alpha_n(G_n - Q_n)$ and $\alpha_n = \frac{1}{n}$, we have

$$nQ_{n+1} = (n-1)Q_n + G_n.$$

By deduction, we get

$$\begin{aligned} nQ_{n+1} &= (n-2)Q_{n-1} + G_{n-1} + G_n \\ &= \{n - [(n-1)+1]\} Q_{n-(n-1)} + G_{n-(n-1)} + G_{n-(n-2)} + \cdots + G_n \\ &= 0 * Q_1 + G_1 + G_2 + \cdots + G_n = \sum_{i=1}^{n} G_i. \end{aligned}$$

As a result, we obtain

$$Q_{n+1} = \frac{1}{n}\sum_{i=1}^{n} G_i, n = 1, 2, \cdots. \tag{1}$$

□

(ii) $\lim_{n\to\infty} V_n = 0$.

Proof. Since $V_n = \mathbb{E}\left[(Q_n - r)^2\right]$, we have

$$\begin{aligned} \lim_{n\to\infty} V_n &= \lim_{n\to\infty} \mathbb{E}\left[(Q_n - r)^2\right] = \lim_{n\to\infty} \mathbb{E}\left[Q_n^2 - 2Q_n r + r^2\right] \\ &= \lim_{n\to\infty} \left\{\mathbb{E}\left[Q_n^2\right] - 2r\mathbb{E}\left[Q_n\right] + r^2\right\} \\ &= \lim_{n\to\infty} \left\{Var\left[Q_n\right] + \mathbb{E}^2\left[Q_n\right] - 2r\mathbb{E}\left[Q_n\right] + r^2\right\}. \end{aligned} \tag{2}$$

Using the result in Eq. 1, i.e., $Q_n = \frac{1}{n-1}\sum_{i=1}^{n-1} G_i$, and given $\mathbb{E}[G_n] = r$, $n = 1, 2 \cdots$, we get

$$\begin{cases} \mathbb{E}\left[Q_n\right] = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n-1} G_i\right] = r \\ Var\left[Q_n\right] = Var\left[\frac{1}{n-1}\sum_{i=1}^{n-1} G_i\right] = \frac{1}{(n-1)^2}\sum_{i=1}^{n-1} Var\left[G_i\right]. \end{cases}$$

Replace this system equations to Eq. 2, we obtain

$$\begin{aligned}\lim_{n\to\infty} V_n &= \lim_{n\to\infty} \left\{ Var\left[Q_n\right] + \mathbb{E}^2\left[Q_n\right] - 2r\mathbb{E}\left[Q_n\right] + r^2 \right\} \\ &= \lim_{n\to\infty} \left\{ \frac{1}{(n-1)^2} \sum_{i=1}^{n-1} Var\left[G_i\right] + r^2 - 2r^2 + r^2 \right\} = 0.\end{aligned}$$

□

(b) (Constant step size) Let $\alpha_n = \alpha$, $0 < \alpha < 2$, show that
(i) $V_{n+1} = (1-\alpha)^2 V_n + \alpha^2 Var[G_n]$, where $Var[G_n] = \mathbb{E}[(G_n - r)^2]$.

Proof. Since $V_n = \mathbb{E}\left[(Q_n - r)^2\right]$, by deduction, we have

$$V_{n+1} = \mathbb{E}\left[(Q_{n+1} - r)^2\right].$$

Due to $\alpha_n = \alpha$, $0 < \alpha < 2$, i.e., $Q_{n+1} = Q_n + \alpha(G_n - Q_n)$, we get

$$\begin{aligned}V_{n+1} &= \mathbb{E}\left[(Q_{n+1} - r)^2\right] = \mathbb{E}\left[\left[Q_n + \alpha\left(G_n - Q_n\right) - r\right]^2\right] = \mathbb{E}\left[\left[(1-\alpha)\, Q_n + \alpha G_n - r\right]^2\right] \\ &= \mathbb{E}\left[\left[(1-\alpha)\, Q_n + \alpha G_n - r + \alpha r - \alpha r\right]^2\right] = \mathbb{E}\left[\left[(1-\alpha)(Q_n - r) + \alpha(G_n - r)\right]^2\right] \\ &= (1-\alpha)^2\, \mathbb{E}\left[(Q_n - r)^2\right] + (1-\alpha)\alpha\mathbb{E}\left[(Q_n - r)(G_n - r)\right] + \alpha^2 \mathbb{E}\left[(G_n - r)^2\right] \\ &= (1-\alpha)^2\, V_n + (1-\alpha)\alpha\mathbb{E}\left[(Q_n - r)(G_n - r)\right] + \alpha^2 Var[G_n].\end{aligned}$$

Consider the term $\mathbb{E}\left[(Q_n - r)(G_n - r)\right]$, by $(Q_n - r) \perp\!\!\!\perp (G_n - r)$ and $\mathbb{E}\left[G_n\right] = r$, $n = 1, 2\cdots$, yielding

$$\mathbb{E}\left[(Q_n - r)(G_n - r)\right] = \mathbb{E}\left[Q_n - r\right]\mathbb{E}\left[G_n - r\right] = \mathbb{E}\left[Q_n - r\right]\left(\mathbb{E}\left[G_n\right] - r\right) = 0.$$

As a result, we obtain

$$V_{n+1} = (1-\alpha)^2 V_n + \alpha^2 Var[G_n], \tag{3}$$

where $Var[G_n] = \mathbb{E}[(G_n - r)^2]$. □

(ii) $\lim_{n\to\infty} \left| V_{n+1} - \frac{\alpha}{2-\alpha} Var[G_n] \right| = 0.$

Proof. Using the result from Eq. 3 and by deduction, we have

$$\begin{aligned}\lim_{n\to\infty} \left| V_{n+1} - \frac{\alpha}{2-\alpha} Var[G_n] \right| &= \lim_{n\to\infty} \left| (1-\alpha)^2 V_n + \alpha^2 Var[G_n] - \frac{\alpha}{2-\alpha} Var[G_n] \right| \\ &= \lim_{n\to\infty} \left| (1-\alpha)^2 \left[(1-\alpha)^2 V_{n-1} + \alpha^2 Var[G_{n-1}]\right] + \alpha^2 Var[G_n] - \frac{\alpha}{2-\alpha} Var[G_n] \right|.\end{aligned}$$

Since G_i are independent and identically distributed random variables $\forall i$, $Var[G_1] = \cdots = Var[G_n]$ holds, getting

$$\begin{aligned}\lim_{n\to\infty} \left| V_{n+1} - \frac{\alpha}{2-\alpha} Var[G_n] \right| &= \lim_{n\to\infty} \left| (1-\alpha)^2 \left[(1-\alpha)^2 V_{n-1} + \alpha^2 Var[G_n]\right] + \alpha^2 Var[G_n] - \frac{\alpha}{2-\alpha} Var[G_n] \right| \\ &= \lim_{n\to\infty} \left| (1-\alpha)^{2n} V_1 + \alpha^2 \left[\sum_{i=1}^{n} (1-\alpha)^{2(n-i)}\, Var[G_n] \right] - \frac{\alpha}{2-\alpha} Var[G_n] \right|.\end{aligned}$$

Applying the triangle inequality and due to $(1-\alpha)^2 < 1$ holds for $0 < \alpha < 2$, we obtain

$$\begin{aligned}\lim_{n\to\infty} \left| V_{n+1} - \frac{\alpha}{2-\alpha} Var[G_n] \right| &\le \lim_{n\to\infty} \left| (1-\alpha)^{2n} V_1 \right| + \lim_{n\to\infty} \left| \alpha^2 \frac{1 - (1-\alpha)^{2n}}{1 - (1-\alpha)^2} Var[G_n] - \frac{\alpha}{2-\alpha} Var[G_n] \right| \\ &= 0 + \frac{\alpha^2}{\alpha(2-\alpha)} Var[G_n] - \frac{\alpha}{2-\alpha} Var[G_n] = 0.\end{aligned}$$

□

2 Problem 2

Recall that the ϵ-greedy algorithm do the following:

$$A_t = \begin{cases} \arg\max_a Q_t(a) & \text{with probability } 1-\epsilon \\ \text{choose } a \text{ uniformly at random} & \text{with probability } \epsilon \end{cases}$$

where $Q_t(a) = \frac{1}{N_t(a)} \sum_{k=1}^{t} R_k \mathbb{I}_{A_k=a}$, $N_t(a) = \sum_{k=1}^{t} \mathbb{I}_{A_k=a}$.
Let $\mathbb{E}[R_t|A_t = a] = q_a$, and $a^* = \arg\max_a q_a$.

(a) Explain briefly why $\epsilon = 0$ is worse than $\epsilon > 0$ in the long run.

When $\epsilon = 0$, then $A_t = \arg\max_a Q_t(a)$, $\forall t$, i.e., there is always exploitation and no exploration. If the arm (action) that the algorithm always exploits is not the global optimal in the long run, then the algorithm is stuck at this sub-optimal arm. As a result, it will be worse than $\epsilon > 0$ because $\epsilon > 0$ means the algorithm can explore other arms before converging to optimal when $t \to \infty$.

(b) Show that

$$\mathbb{E}[R_t] = \sum_a q_a \mathbb{P}(A_t = a),$$

(If necessary, you may assume that R_t is a discrete random variable taking finite many values).

Proof. We have R_t is a random variable whose expected value $\mathbb{E}[R_T]$ is defined, and A_t is also a random variable on the finite action space $\mathcal{A}$, then by definition of the Law of total expectation, we have

$$\mathbb{E}[R_t] = \mathbb{E}_{A_t \sim \mathcal{A}}[\mathbb{E}[R_t|A_t]] = \sum_a \mathbb{E}[R_t|A_t = a]\mathbb{P}(A_t = a).$$

Combining with the notation $\mathbb{E}[R_t|A_t = a] = q_a$, we obtain

$$\mathbb{E}[R_t] = \sum_a \mathbb{E}[R_t|A_t = a]\mathbb{P}(A_t = a) = \sum_a q_a \mathbb{P}(A_t = a). \tag{4}$$

□

(c) With $\epsilon > 0$, suppose at time T we have

$$\arg\max_a Q_T(a) = a^*,$$

compute $\mathbb{E}[R_T]$. How does $\mathbb{E}[R_T]$ changes as we increase ϵ, the exploration rate?

By definition of ϵ-greedy algorithm, we have A_t follows a Bernoulli distribution with parameter ϵ, i.e., $A_t = \arg\max_a Q_t(a)$ with probability $1-\epsilon$ and A_t is chosen a uniformly at random over n possible action in set $\mathcal{A}$, with probability ϵ. So, since $\arg\max_a Q_T(a) = a^*$, from Eq. 4, we have

$$\mathbb{E}[R_T] = \sum_a q_a \mathbb{P}(A_T = a) = (1-\epsilon)q_{a^*} + \frac{\epsilon}{n}\sum_a q_a = q_{a^*} - \epsilon\left(q_{a^*} - \frac{1}{n}\sum_a q_a\right). \tag{5}$$

On the other hand, by $\arg\max_a Q_T(a) = a^*$, then $q_{a^*} \geq q_a$ holds $\forall a \in \mathcal{A}$, yielding

$$nq_{a^*} \geq \sum_a q_a \Rightarrow q_{a^*} - \frac{1}{n}\sum_a q_a \geq 0.$$

Apply this result to Eq. 5, for positive $\epsilon_1 < \epsilon_2$, we obtain

$$q_{a^*} - \epsilon_1\left(q_{a^*} - \frac{1}{n}\sum_a q_a\right) \geq q_{a^*} - \epsilon_2\left(q_{a^*} - \frac{1}{n}\sum_a q_a\right),$$

i.e., $\mathbb{E}[R_T]$ non-increase or decrease as we increase ϵ.

(d) In the comparison shown in Fig. 2.2, which method will perform best in the long run in terms of cumulative reward and probability of selecting the best action? How much better will it be? Express your answer quantitatively.

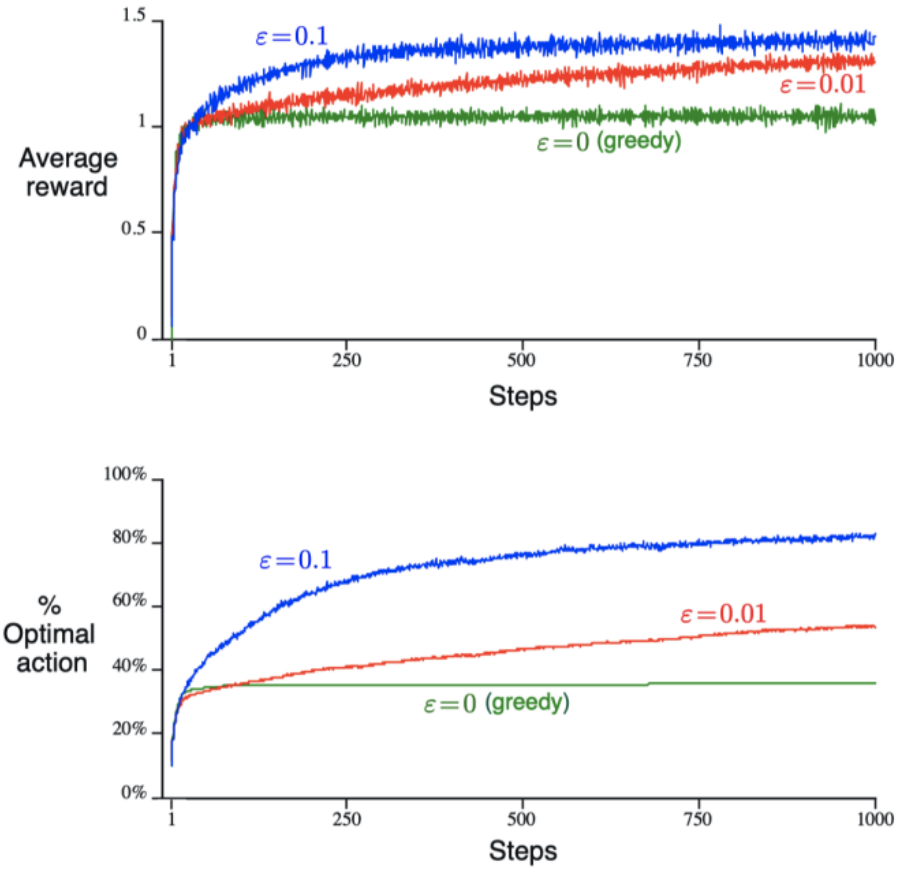

Figure 2.2: Average performance of ε-greedy action-value methods on the 10-armed testbed. These data are averages over 2000 runs with different bandit problems. All methods used sample averages as their action-value estimates.

Using the result in question (a), we know that $\epsilon = 0$ is worse than $\epsilon > 0$ in the long run. And $\epsilon = 0.01$ is the best setting in the long run because of the following reasons:
Consider the average reward in Fig. 2.2, by definition, we have the cumulative reward function w.r.t. ϵ is

$$f(\epsilon) = \lim_{T\to\infty} \frac{\mathbb{E}[\sum_{t=0}^{T} R_t]}{T} = (1 - \epsilon + \frac{\epsilon}{n})q_{a^*} + \frac{\epsilon}{n}\sum_{a\neq a^*} q_a.$$

Compare the cumulative reward between $\epsilon = 0.01$ and $\epsilon = 0.1$, with $n = 10$ (in Fig), we have

$$f(0.01) - f(0.1) = 0.991q_{a^*} + 0.001\sum_{a\neq a^*} q_a - 0.91q_{a^*} - 0.01\sum_{a\neq a^*} q_a = 0.081q_{a^*} - 0.009\sum_{a\neq a^*} q_a.$$

So, we obtain $\epsilon = 0.01$ is better than $\epsilon = 0.1$ by $0.081q_{a^*} - 0.009\sum_{a\neq a^*} q_a$ in terms of cumulative reward.
Similarly, consider the optimal action in Fig. 2.2, by definition, we have the probability of selecting the best action function w.r.t. ϵ is

$$g(\epsilon) = 1 - \epsilon + \frac{\epsilon}{n}.$$

Compare the probability of selecting the best action between $\epsilon = 0.01$ and $\epsilon = 0.1$, with $n = 10$ (in Fig), we have

$$g(0.01) - g(0.1) = 1 - 0.01 + \frac{0.01}{10} - 1 + 0.1 - \frac{0.1}{10} = 0.081.$$

So, we obtain $\epsilon = 0.01$ is better than $\epsilon = 0.1$ by 8.1% in terms of the probability of selecting the best action.