Fakultät für Wirtschaftsinformatik
und Wirtschaftsmathematik

Universität Mannheim

Prof. Dr. Leif Döring
Sara Klein, Benedikt Wille

Reinforcement Learning

1. Exercise Sheet - Solutions

1. The Regret

Recall Definition 1.1.6 from the lecture. Suppose ν is a bandit model and $(\pi_t)_{t=1,\dots,n}$ a learning strategy. Then the regret is defined by

$$R_n(\pi) := nQ_* - \mathbb{E}_\pi\Big[\sum_{t=1}^n X_t\Big], \quad n \in \mathbb{N},$$

where $Q_* := \int_{-\infty}^{\infty} x P_{a_*}(dx)$ the expected reward of the best arm $a_* = \mathrm{argmax}_a Q_a$.

a) Suppose a two-armed bandit with $Q_1 = 1$ and $Q_2 = -1$ and a learning strategy π given by

$$\pi_t = \begin{cases} \delta_1, & t \text{ even,} \\ \delta_2, & t \text{ odd.} \end{cases}$$

Calculate the regret $R_n(\pi)$ for all $n \in \mathbb{N}$.

Solution:
If $n \in \mathbb{N}$ is even, then

$$R_n(\pi) = nQ_* - \mathbb{E}^\pi\Big[\sum_{t\le n} X_t\Big] = n * 1 - \big(\frac{n}{2}(-1) - \frac{n}{2}1\big) = n \tag{1}$$

and if $n \in \mathbb{N}$ is odd, then

$$\begin{aligned} R_n(\pi) =& nQ_* - \mathbb{E}^\pi\Big[\sum_{t\le n} X_t\Big] \\ &= (n-1)Q_* - \mathbb{E}^\pi\Big[\sum_{t\le n-1} X_t\Big] + Q_* - \mathbb{E}^\pi[X_n] \\ &= R_{n-1}(\pi) + 1 - (-1) \\ &\overset{(1)}{=} n - 1 + 1 + 1 = n + 1 \end{aligned}$$

b) Define a stochastic bandit and a learning strategy such that the regret is 5 for all $n \ge 5$.

Solution:
Consider for example the 3*-armed bandit with* $Q_1 = 1, Q_2 = -1, Q_3 = 0$ *and a policy* π *with*

$$\pi_1 = \pi_2 = \delta_2, \quad \pi_3 = \delta_3, \quad \pi_t = \delta_1 \, \forall t \ge 4.$$

Then for all $n \geq 4$ we have

$$\begin{aligned} R_n(\pi) &= nQ_* - \mathbb{E}^\pi\Big[\sum_{t\leq n} X_t\Big] \\ &= n * 1 - \Big((-1) + (-1) + 0 + \sum_{t=4}^{n} 1\Big) = n + 2 - (n-3) = 5. \end{aligned}$$

c) Show for all learning strategies π that $R_n(\pi) \geq 0$ and $\limsup_{n\to\infty} \frac{R_n(\pi)}{n} < \infty$.

Solution:

Claim: for all learning strategies π that $R_n(\pi) \geq 0$ and $\limsup_{n\to\infty} \frac{R_n(\pi)}{n} < \infty$.

Proof: Fix a learning strategy π. Then for the first Claim

$$\begin{aligned} R_n(\pi) &= nQ_* - \mathbb{E}^\pi\Big[\sum_{t\leq n} X_t\Big] \\ &= nQ_* - \sum_{t\leq n} \mathbb{E}^\pi\Big[X_t\Big] \\ &= nQ_* - \sum_{t\leq n}\sum_{a\in\mathcal{A}} \mathbb{E}^\pi\Big[X_t \mathbf{1}_{\{A_t=a\}}\Big] \\ &= nQ_* - \sum_{t\leq n}\sum_{a\in\mathcal{A}} \mathbb{P}^\pi(A_t=a)\mathbb{E}^\pi\Big[X_t\Big|A_t=a\Big] \\ &= nQ_* - \sum_{t\leq n}\sum_{a\in\mathcal{A}} \mathbb{P}^\pi(A_t=a)Q_a \\ &\geq nQ_* - \sum_{t\leq n}\sum_{a\in\mathcal{A}} \mathbb{P}^\pi(A_t=a)Q_* \\ &= nQ_* - nQ_* \\ &= 0, \end{aligned}$$

where we used the formular for conditional expectation in the forth line, the definition of Q_a in the fifth line and $Q_a \leq Q_*$ for all $a \in \mathcal{A}$ in the inequality.

For the second Claim we define $Q_{-*} := \min_{a\in\mathcal{A}} Q_a$. Then it holds similar to the calculation above

$$\begin{aligned} R_n(\pi) &= nQ_* - \sum_{t\leq n}\sum_{a\in\mathcal{A}} \mathbb{P}^\pi(A_t=a)Q_a \\ &\leq nQ_* - \sum_{t\leq n}\sum_{a\in\mathcal{A}} \mathbb{P}^\pi(A_t=a)Q_{-*} \\ &= nQ_* - nQ_{-*}. \end{aligned}$$

Thus

$$\limsup_{n\to\infty} \frac{R_n(\pi)}{n} \leq \limsup_{n\to\infty} \frac{nQ_* - nQ_{-*}}{n} = Q_* - Q_{-*} < \infty.$$

d) Let $R_n(\pi) = 0$. Prove that π is deterministic, i.e. all π_t are almost surely constant and only chose the best arm.

Solution:

Claim: If $R_n(\pi) = 0$ for all $n \geq 1$, then π is deterministic and $\pi_t = \delta_{a^*}$ almost surely.

Proof: Let $R_n(\pi) = 0$ for all $n \geq 1$ and assume there exists $t \geq 1$ such that $\pi_t \neq \delta_{a^*}$. Then there exists an arm $a \neq a^*$ with $Q_a < Q_{a^*}$ such that $\mathbb{P}^\pi(A_t = a) > 0$. We follow

$$\begin{aligned}
\mathbb{E}^\pi[X_t] &= \sum_{a' \in \mathcal{A}} \mathbb{P}^\pi(A_t = a')Q_{a'} \\
&= \mathbb{P}^\pi(A_t = a)Q_a + \sum_{a' \neq a} \mathbb{P}^\pi(A_t = a')Q_{a'} \\
&\leq \mathbb{P}^\pi(A_t = a)Q_a + (1 - \mathbb{P}^\pi(A_t = a))Q_* \\
&= Q_* + \mathbb{P}^\pi(A_t = a)(Q_a - Q_*) \\
&< Q_*.
\end{aligned}$$

Using this we have for all $n \geq t$

$$\begin{aligned}
R_n(\pi) &= nQ_* - \sum_{t \leq n} \mathbb{E}^\pi\Big[X_t\Big] \\
&\geq nQ_* - \Big((n-1)Q_* + \mathbb{E}^\pi[X_t]\Big) \\
&> Q_* - Q_* = 0.
\end{aligned}$$

This is a contradiction.

e) Suppose ν is a 1-subgaussian bandit model with k arms and $km \leq n$, then consider the Explore then Commit algorithm and recall the regret bound:

$$R_n \leq \underbrace{m \sum_{a \in \mathcal{A}} \Delta_a}_{\text{exploration}} + \underbrace{(n - mk) \sum_{a \in \mathcal{A}} \Delta_a \exp\Big(-\frac{m\Delta_a^2}{4}\Big)}_{\text{exploitation}}.$$

Assume now $k = 2$, such that $\Delta_1 = 0$ and $\Delta_2 = \Delta$ then we get

$$R_n \leq m\Delta + (n - m2)\Delta \exp\Big(-\frac{m\Delta^2}{4}\Big) \leq m\Delta + n\Delta \exp\Big(-\frac{m\Delta^2}{4}\Big).$$

Show that this upper bound is minimized for $m = \max\Big\{1, \Big\lceil \frac{4}{\Delta^2}\log(\frac{n\Delta^2}{4})\Big\rceil\Big\}$.

Solution:

Define the function $f(m) = m\Delta + n\Delta\exp\Big(-\frac{m\Delta^2}{4}\Big)$ with $n > 0, \Delta > 0$. First show that f is convex, then we can solve for a minimum in $\mathbb{R}$ to find minimizers in the natural numbers. Note therefore that

$$\begin{aligned}
\nabla f(m) &= \Delta - \frac{n\Delta^3}{4}\exp\Big(-\frac{m\Delta^2}{4}\Big) \\
\nabla^2 f(m) &= \frac{n\Delta^5}{16}\exp\Big(-\frac{m\Delta^2}{4}\Big) > 0 \quad \forall m \in \mathbb{R}.
\end{aligned}$$

Solving $\nabla f(m) = 0$ yields

$$\begin{aligned} \frac{n\Delta^3}{4} \exp\Big(- \frac{m\Delta^2}{4}\Big) &= \Delta \\ \Leftrightarrow \quad m &= \frac{4}{\Delta^2} \log\Big(\frac{n\Delta^2}{4}\Big). \end{aligned}$$

Defining our candidate $m^* = \frac{4}{\Delta^2}\log\Big(\frac{n\Delta^2}{4}\Big)$ we conclude from

$$\nabla^3 f(m) = -\frac{n\Delta^7}{64} \exp\Big(- \frac{m\Delta^2}{4}\Big) < 0$$

that f increases to the left of m^* faster than to the right, such that $f(\Big\lceil \frac{4}{\Delta^2}\log(\frac{n\Delta^2}{4})\Big\rceil) < f(\Big\lfloor \frac{4}{\Delta^2}\log(\frac{n\Delta^2}{4})\Big\rfloor)$. As m has to be a natural number we know $m \geq 1$ and so

$$m = \max\Big\{1, \Big\lceil \frac{4}{\Delta^2}\log(\frac{n\Delta^2}{4})\Big\rceil\Big\}$$

minimizes the regret.

2. The Regret - Part 2

Show the following two claims.

a) If the failure probabilities do not decay to zero then the regret grows linearly.
Solution:
By Lemma 1.2.10 in the lecture notes we know that

$$R_n(\pi) \geq \min_{a \neq a^*} \Delta_a \sum_{t=1}^{n} \tau_t(\pi).$$

Assume now that the failure probabilities do not decay to zero, i.e. there exist $c > 0$ and $T \geq 1$ such that $\tau_t(\pi) > c$ for all $t \geq T$. Then for all $n > T$ we have

$$\begin{aligned} R_n(\pi) &\geq \min_{a \neq a^*} \Delta_a \left(\sum_{t=1}^{T} \tau_t(\pi) + (n-T)c \right) \\ &\geq (n-T)c \min_{a \neq a^*} \Delta_a. \end{aligned}$$

Thus, we have shown that the regret grows at least linearly in n for n large enough.
To see that the regret also grows at most linearly in n, note that

$$R_n(\pi) \leq \max_{a \in \mathcal{A}} \Delta_a \sum_{t=1}^{n} \tau_t(\pi)$$

$$\leq n \max_{a \in \mathcal{A}}.$$

This proves the claim.

b) If the failure probability $\tau_n(\pi)$ behaves like $\frac{1}{n}$, then the regret behaves like $\sum_{a\in\mathcal{A}} \Delta_a \log(n)$ with constants that depend on the concrete bandit model.

Hint: Recall from basic analysis that $\int_1^n \frac{1}{x}dx = \log(n)$ and how to relate sums and integrals for monotone integrands.

Solution:

Again by Lemma 1.2.10 in the lecture notes we know that

$$R_n \leq \max_{a\in\mathcal{A}} \Delta_a \sum_{t=1}^{n} \tau_t(\pi) \quad \textit{and} \quad R_n(\pi) \geq \min_{a\neq a^*} \Delta_a \sum_{t=1}^{n} \tau_t(\pi).$$

For $\tau_n(\pi) \simeq \frac{1}{n}$ *we will prove that* $\log(n) \leq \sum_{t=1}^{n} \frac{1}{t} \leq \log(n) + 1$.

First recall that $I = \{t\}_{t=1}^{n}$ *can be interpreted as a disjoint decomposition of the interval* $[1, n]$ *each of length* 1. *Next, we upper and lower bound the integral* $\int_1^t \frac{1}{x}dx$ *by taking into accout that* $\frac{1}{x}$ *is monotonic decreasing and considering the upper-sum and lower-sum. We obtain*

$$\sum_{t=2}^{n} \frac{1}{t} \leq \int_1^t \frac{1}{x}dx \leq \sum_{t=1}^{n-1} \frac{1}{t}.$$

Thus, we follow that

$$\sum_{t=1}^{n} \frac{1}{t} \geq \sum_{t=1}^{n-1} \frac{1}{t} \geq \log(n)$$

and on the other hand

$$\sum_{t=1}^{n} \frac{1}{t} = 1 + \sum_{t=2}^{n} \frac{1}{t} \leq 1 + \log(n).$$

All in all we see that

$$R_n \leq \max_{a\in\mathcal{A}} \Delta_a \sum_{t=1}^{n} \tau_t(\pi) \leq \max_{a\in\mathcal{A}} \Delta_a (1 + \log(n))$$

and

$$R_n(\pi) \geq \min_{a\neq a^*} \Delta_a \sum_{t=1}^{n} \tau_t(\pi) \geq \min_{a\neq a^*} \Delta_a \log(n).$$

We conclude the claim by realizing that $\min_{a\neq a^*} \Delta_a \leq \sum_a \Delta_a \leq K \max_a \Delta_a$, *where* K *is the number of arms. Hence, there exists a constant* $\tilde{C}$ *(dependent on the* Δ_a*'s) such that* $R_n = \tilde{C} \sum_{a\in\mathcal{A}} \Delta_a \log(n)$.