FAKULTÄT FÜR WIRTSCHAFTSINFORMATIK
UND WIRTSCHAFTSMATHEMATIK

UNIVERSITÄT MANNHEIM

Prof. Dr. Leif Döring
André Ferdinand, Sara Klein

Reinforcement Learning

1. Excercise 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 $Q_* := \int_{-\infty}^{\infty} xP_{a_*}(dx)$ the expected reward of the best arm $a_* = \text{argmax}_a Q_a$.

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 exits $t \geq 1$ such that $\pi_t \neq \delta_{a^*}$. Then there exits 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 explorte 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 minimisers 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*

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

As m *has to be a natural number we know* $m \geq 1$ *and so*

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

minimises the regret.