Prof. Dr. Leif Döring Benedikt Wille

1. Exercise Sheet - Solutions

Reinforcement Learning 25.02.2025

UNIVERSITÄT Mannheim

1. The Regret - Part 1

Recall Definitions 1.2.1 and 1.2.6 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

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

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

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

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

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

$$R_{n}(\pi) = nQ_{*} - \mathbb{E}^{\pi} \left[\sum_{t \leq n} X_{t} \right]$$

= $(n-1)Q_{*} - \mathbb{E}^{\pi} \left[\sum_{t \leq n-1} X_{t} \right] + Q_{*} - \mathbb{E}^{\pi} [X_{n}]$
= $R_{n-1}(\pi) + 1 - (-1)$
 $\stackrel{(1)}{=} n - 1 + 1 + 1 = n + 1$

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 $n \ge 4$ we have

$$R_n(\pi) = nQ_* - \mathbb{E}^{\pi} \left[\sum_{t \le n} X_t \right]$$
$$= n * 1 - \left((-1) + (-1) + 0 + \sum_{t=4}^n 1 \right) = n + 2 - (n - 3) = 5.$$

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

Claim: for all learning strategies π that $R_n(\pi) \ge 0$ and $\limsup_{n\to\infty} \frac{R_n(\pi)}{n} < \infty$. Proof: Fix a learning strategy π . Then for the first Claim

$$R_{n}(\pi) = nQ_{*} - \mathbb{E}^{\pi} \left[\sum_{t \leq n} X_{t} \right]$$
$$= nQ_{*} - \sum_{t \leq n} \mathbb{E}^{\pi} \left[X_{t} \right]$$
$$= nQ_{*} - \sum_{t \leq n} \sum_{a \in \mathcal{A}} \mathbb{E}^{\pi} \left[X_{t} \mathbf{1}_{\{A_{t}=a\}} \right]$$
$$= nQ_{*} - \sum_{t \leq n} \sum_{a \in \mathcal{A}} \mathbb{P}^{\pi} (A_{t} = a) \mathbb{E}^{\pi} \left[X_{t} \middle| A_{t} = a \right]$$
$$= nQ_{*} - \sum_{t \leq n} \sum_{a \in \mathcal{A}} \mathbb{P}^{\pi} (A_{t} = a)Q_{a}$$
$$\geq nQ_{*} - \sum_{t \leq n} \sum_{a \in \mathcal{A}} \mathbb{P}^{\pi} (A_{t} = a)Q_{*}$$
$$= nQ_{*} - nQ_{*}$$
$$= 0,$$

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

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

Thus

above

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

d) Let $R_n(\pi) = 0$. Suppose that the best arm is unique. 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 $n \ge 1$, then π is deterministic and $\pi_t = \delta_{a^*}$ almost surely. Proof: Let $R_n(\pi) = 0$ for all $n \ge 1$ and assume there exists $t \ge 1$ such that $\pi_t \ne \delta_{a^*}$. Then there exists an arm $a \neq a^*$ with $Q_a < Q_{a^*}$ such that $\mathbb{P}^{\pi}(A_t = a) > 0$. We follow

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

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

$$R_n(\pi) = nQ_* - \sum_{t \le n} \mathbb{E}^{\pi} \left[X_t \right]$$

$$\geq nQ_* - \left((n-1)Q_* + \mathbb{E}^{\pi} [X_t] \right)$$

$$> Q_* - Q_* = 0.$$

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:

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

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

$$R_n \le m\Delta + (n-m2)\Delta \exp\left(-\frac{m\Delta^2}{4}\right) \le m\Delta + n\Delta \exp\left(-\frac{m\Delta^2}{4}\right).$$

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

Define the function $f(m) = m\Delta + n\Delta \exp\left(-\frac{m\Delta^2}{4}\right)$ 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

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

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

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

Defining our candidate $m^* = \frac{4}{\Delta^2} \log\left(\frac{n\Delta^2}{4}\right)$ 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 $f(\left\lceil \frac{4}{\Delta^2} \log(\frac{n\Delta^2}{4}) \right\rceil) < f(\left\lfloor \frac{4}{\Delta^2} \log(\frac{n\Delta^2}{4}) \right\rfloor)$. As m has to be a natural number we know $m \ge 1$ and so

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

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

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

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

$$R_n(\pi) \ge \min_{a \ne a^*} \Delta_a \left(\sum_{t=1}^T \tau_t(\pi) + (n-T)c \right)$$
$$\ge (n-T)c \min_{a \ne a^*} \Delta_a.$$

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

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

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

$$R_n \le \max_{a \in \mathcal{A}} \Delta_a \sum_{t=1}^n \tau_t(\pi) \quad and \quad R_n(\pi) \ge \min_{a \ne a^*} \Delta_a \sum_{t=1}^n \tau_t(\pi).$$

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

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

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

Thus, we follow that

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

and on the other hand

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

All in all we see that

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

and

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

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 $R_n = \tilde{C} \sum_{a \in \mathcal{A}} \Delta_a \log(n)$.

3. Sub-Gaussian random variables

Recall Definition 1.3.3. of a σ -sub-Gaussian random variable X.

a) Show that every σ -sub-Gaussian random variable satisfies $\mathbb{E}[X] = 0$ and $\mathbb{V}[X] \leq \sigma^2$. Solution:

Let X be a σ -sub-Gaussian random variable. Then by Fubini

$$\sum_{t\geq 0} \frac{\lambda^t}{t!} \mathbb{E}[X^t] = \mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\lambda^2 \sigma^2}{2}} = \sum_{t\geq 0} \frac{\lambda^{2t} \sigma^{2t}}{2^t t!}.$$
(2)

We follow that

$$\lambda \mathbb{E}[X] + \frac{\lambda^2}{2} \mathbb{E}[X^2] \le \frac{\lambda^2 \sigma^2}{2} + g(\lambda), \tag{3}$$

for

$$g(\lambda) = \sum_{t \ge 2} \frac{\lambda^{2t} \sigma^{2t}}{2^t t!} - \sum_{t \ge 3} \frac{\lambda^t}{t!} \mathbb{E}[X^t]$$

Note that $g \in o(\lambda^2)$ because

$$\lim_{\lambda \to 0} \frac{g(\lambda)}{\lambda^2} = \sum_{t \ge 2} \lim_{\lambda \to 0} \frac{\lambda^{2t} \sigma^{2t}}{2^t t!} - \sum_{t \ge 3} \lim_{\lambda \to 0} \frac{\lambda^t}{t!} \mathbb{E}[X^t] = 0,$$

where we used that both sums are finite due to the finiteness of exp. Finally, for $\lambda > 0$ dividing (3) by $1/\lambda$ and taking the limits $\lambda \downarrow 0$ leads to

$$\mathbb{E}[X] \le \frac{\lambda \sigma^2}{2} + \frac{g(\lambda)}{\lambda} - \frac{\lambda}{2} \mathbb{E}[X^2] \to 0, \quad \lambda \downarrow 0$$

and for $\lambda < 0$ similarly

$$\mathbb{E}[X] \ge \frac{\lambda \sigma^2}{2} + \frac{g(\lambda)}{\lambda} - \frac{\lambda}{2} \mathbb{E}[X^2] \to 0, \quad \lambda \uparrow 0.$$

Hence, $\mathbb{E}[X] = 0$.

Rewriting (3) once again and deviding by λ^2 results in

$$\mathbb{E}[X^2] \le 2\left(\frac{\sigma^2}{2} + \frac{g(\lambda)}{\lambda^2}\right) \to \sigma^2, \quad \lambda \to 0,$$

which proofs the second claim.

b) Suppose X is σ -sub-Gaussian. Prove that cX is $|c|\sigma$ -sub-Gaussian. Solution: We have

$$M_{cX}(\lambda) = \mathbb{E}\left[e^{\lambda cX}\right] \le e^{\frac{(c\lambda)^2 \sigma^2}{2}} = e^{\frac{\lambda^2 (c\sigma)^2}{2}}.$$

Thus, cX is $|c|\sigma$ -sub-Gaussian.

c) Show that $X_1 + X_2$ is $\sqrt{\sigma_1^2 + \sigma_2^2}$ -sub-Gaussian if X_1 and X_2 are independent σ_1 -sub-Gaussian and σ_2 -sub-Gaussian random variables. Solution:

We have

$$M_{X_1+X_2}(\lambda) = \mathbb{E}\left[e^{\lambda(X_1+X_2)}\right] = \mathbb{E}\left[e^{\lambda X_1}e^{\lambda X_2}\right]$$
$$= \mathbb{E}\left[e^{\lambda X_1}\right]\mathbb{E}\left[e^{\lambda X_2}\right]$$
$$\leq e^{\frac{\lambda^2 \sigma_1^2}{2}}e^{\frac{\lambda^2 \sigma_2^2}{2}}$$
$$= \exp(\frac{\lambda^2(\sqrt{\sigma_1^2 + \sigma_2^2})^2}{2}).$$

where the thrid equality follows from independence. This proofs the claim.

d) Show that a Bernoulli-variable is $\frac{1}{2}$ -sub-Gaussian. Solution:

Exactly as in the next exercise but with a = 0 and b = 1.

e) Show that every centered bounded random variable, say bounded below by a and above by b is $\frac{(b-a)}{2}$ -sub-Gaussian.

Solution:

As $a \leq X \leq b$ we have almost surely

$$e^{\lambda X} \le \frac{b-X}{b-a}e^{\lambda a} + \frac{X-a}{b-a}e^{\lambda b}.$$

We follow

$$\begin{split} \mathbf{E}\Big[e^{\lambda X}\Big] &\leq \frac{b - \mathbf{E}[X]}{b - a} e^{\lambda a} + \frac{\mathbf{E}[X] - a}{b - a} e^{\lambda b} \\ &= \frac{b}{b - a} e^{\lambda a} - \frac{a}{b - a} e^{\lambda b} \\ &= \exp L(\lambda(b - a)), \end{split}$$

where we used $\mathbb{E}[X] = 0$ and L(h) is definied by

$$L(h) = \frac{ha}{(b-a)} + \log\left(1 + \frac{a - e^h a}{b-a}\right).$$

We will show that $L(h) \leq h^2/8$, then it follows

$$\mathbb{E}\left[e^{\lambda X}\right] \le \exp L(\lambda(b-a)) \le \exp(\frac{\lambda^2(b-a)^2}{8}),$$

which proofs that X is σ -sub-Gauss with $\sigma = \frac{(b-a)}{2}$. So let us proof that $L(h) \leq h^2/8$. Therefore we first calculate the first and second derivative.

$$\begin{split} \nabla L(h) &= \frac{a}{b-a} - \frac{e^h a}{b-e^h a}, \\ \nabla^2 L(h) &= -\frac{e^h a b}{(b-e^h a)^2}. \end{split}$$

Note now, that

$$\begin{split} L(0) &= 0, \\ \nabla L(0) &= 0 \quad and \\ \nabla^2 L(h) &= -\frac{e^h a b}{\underbrace{(b-e^h a)^2}_{\geq -4(be^h a)}} \leq \frac{e^h a b}{4e^h a b} \leq \frac{1}{4}. \end{split}$$

By Taylor we know there exists $\theta \in [0,1]$ such that

$$L(h) = L(0) + h\nabla L(0) + \frac{1}{2}h^2\nabla^2 L(h\theta) = \frac{1}{2}h^2\nabla^2 L(h\theta).$$

As $\nabla^2 L(h) \leq \frac{1}{4}$, we have

$$L(h) \le \frac{1}{2}h^2\frac{1}{4} = \frac{h^2}{8}.$$

This conclues the proof.

4. Regret Bound

Recall the upper bound on the regret for ETC in the case of two arms from exercise 1. Show that

$$R_n(\pi) \le \Delta + C\sqrt{n}$$

for some model-free constant C so that, in particular, $R_n(\pi) \leq 1 + C\sqrt{n}$ for all bandit models with regret bound $\Delta \leq 1$ (for instance for Bernoulli bandits).

Hint: Use the same trick as in the proof of Theorem 1.2.10.

Solution:

We will first show that

$$R_n(\pi) \le \min\{n\Delta, \Delta + \frac{4}{\Delta} \left(1 + \max\{0, \log(\frac{n\Delta^2}{4})\}\right)\}$$

by plugging $m^* = \max\{1, \lceil \frac{4}{\Delta^2} \log(\frac{n\Delta^2}{4}) \rceil\}$ into the regret bound from the last exercise sheet

$$R_n \le m^* \Delta + (n - 2m^*) \Delta \exp(-\frac{m^* \Delta^2}{4})$$

This leads to

$$\begin{split} R_n &\leq m^* \Delta + (n - 2m^*) \Delta \exp(-\frac{\Delta^2}{4} \max\{1, \lceil \frac{4}{\Delta^2} \log(\frac{n\Delta^2}{4}) \rceil\}) \\ &= m^* \Delta + (n - 2m^*) \Delta \min\{\exp(-\frac{\Delta^2}{4}), \underbrace{\exp(-\frac{\Delta^2}{4} \lceil \frac{4}{\Delta^2} \log(\frac{n\Delta^2}{4}) \rceil)\}}_{\leq \exp(-\frac{\Delta^2}{4} \frac{4}{\Delta^2} \log(\frac{n\Delta^2}{4})) \leq \frac{4}{\Delta^2 n}} \\ &\leq m^* \Delta + \min\{(n - 2m^*) \Delta \exp(-\frac{\Delta^2}{4}), (n - \underbrace{2m^*}_{>0}) \Delta \frac{4}{\Delta^2 n}\} \\ &\leq m^* \Delta + \min\{(n - 2m^*) \Delta \exp(-\frac{\Delta^2}{4}), \frac{4}{\Delta}\} \\ &\leq \min\left\{m^* \Delta + (n - 2m^*) \Delta \underbrace{\exp(-\frac{\Delta^2}{4})}_{\leq 1}, \frac{4}{\Delta} + \underbrace{\max\{1, \lceil \frac{4}{\Delta^2} \log(\frac{n\Delta^2}{4}) \rceil\} \Delta}_{\leq (1 + \max\{0, \log(\frac{n\Delta^2}{4})\})}\right\} \\ &\leq \min\left\{\underbrace{-m^* \Delta}_{\leq 0} + n\Delta, \Delta + \frac{4}{\Delta} \left(1 + \max\{0, \log(\frac{n\Delta^2}{4})\}\right)\right\}. \end{split}$$

Using this we can devide in the cases $\Delta \leq \sqrt{\frac{c}{n}}$ and $\Delta > \sqrt{\frac{c}{n}}$, for some constant c > 0 which we specify later. Thus, in the first case $\Delta \leq \sqrt{\frac{c}{n}}$ we have

$$R_n \le \min\left\{n\Delta, \Delta + \frac{4}{\Delta}\left(1 + \max\{0, \log(\frac{n\Delta^2}{4})\}\right)\right\} \le n\Delta \le \sqrt{cn}.$$

For the second case we consider the second term and rewrite

$$\frac{4}{\Delta} \left(1 + \max\{0, \log(\frac{n\Delta^2}{4})\} \right) \le 4\left(\frac{1}{\Delta} + \frac{\log(\frac{n\Delta^2}{4})}{\Delta}\right).$$

We define $f(x) = \frac{\log(\frac{nx^2}{4})}{x}$, and prove $f(x) \le 2$ for $x \ge \sqrt{\frac{e^24}{n}}$. If this is true we have for the second case with $c = e^24$ that

$$R_n \le \Delta + 4\left(\frac{1}{\Delta} + \frac{\log(\frac{n\Delta^2}{4})}{\Delta}\right)$$
$$\le \Delta + 4\left(\sqrt{\frac{n}{c}} + 2\right) \le \Delta + \sqrt{n}\left(8 + \frac{4}{\sqrt{c}}\right) = \Delta + \sqrt{n}\left(8 + \frac{2}{e}\right)$$

Now to our claim. We have

$$f'(x) = \frac{2 - \log(\frac{nx^2}{4})}{x^2}$$

and so $f'(x) \leq 0$ iff

$$\log(\frac{nx^2}{4}) \ge 2 \quad \Leftrightarrow \quad x \ge \sqrt{\frac{e^24}{n}}.$$

Thus f decreases in $\left[\sqrt{\frac{e^24}{n}},\infty\right)$ and so $f(x) \leq f(\sqrt{\frac{e^24}{n}}) = 2$. Coosing $C = 8 + \frac{2}{e}$ concludes the proof, as for the first case with $c = e^24$ we have $R_n \leq 2e\sqrt{n} \leq \Delta + C\sqrt{n}$ and for the second case also $R_n \leq \Delta + C\sqrt{n}$.

5. Advanced: ϵ -greedy Regret

Let π the learning strategy that first explores each arm once and then continuous according to ϵ -greedy for some $\epsilon \in (0, 1)$ fixed. Furthermore, assume that ν is a 1-sub-gaussian bandit model. Show that the regret grows linearly:

$$\lim_{n \to \infty} \frac{R_n(\pi)}{n} = \frac{\epsilon}{K} \sum_{a \in \mathcal{A}} \Delta_a$$

Solution:

We denote by π the learning strategy induced by the ϵ -greedy algorithm. Further, denote by $\hat{Q}_a(t) = \frac{1}{N_a(t)} \sum_{n=0}^t X_n^{\pi} \mathbf{1}_{A_n^{\pi}=a}$ the estimator for arm a after round t. Then, for $n \geq K$

$$\mathbb{P}(A_t^{\pi} = a) = \frac{\epsilon}{K} + (1 - \epsilon)\mathbb{P}(\hat{Q}_a(t) \ge \max_b \hat{Q}_b(t)).$$

By the regret decomposition lemma we follow directly that

$$\lim_{n \to \infty} \frac{R_n(\pi)}{n} = \lim_{n \to \infty} \frac{1}{n} \sum_{a \in \mathcal{A}} \Delta_a \sum_{t=1}^n \mathbb{P}(A_t^{\pi} = a)$$
$$\geq \sum_{a \in \mathcal{A}} \Delta_a \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^n \frac{\epsilon}{K}$$
$$= \frac{\epsilon}{K} \sum_{a \in \mathcal{A}} \Delta_a.$$

To show the upper bound we will prove that $\sum_{t=1}^{\infty} \mathbb{P}(\hat{Q}_a(t) \ge \max_b \hat{Q}_b(t)) \le C < \infty$. Then the claim follows again from the regret decomposition lemma:

$$\lim_{n \to \infty} \frac{R_n(\pi)}{n} = \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^n \sum_{a \in \mathcal{A}} \Delta_a \mathbb{P}(A_t^{\pi} = a)$$
$$= \lim_{n \to \infty} \sum_{a \in \mathcal{A}} \Delta_a \frac{1}{n} \sum_{t=1}^n \left(\frac{\epsilon}{K} + P(\hat{Q}_a(t) \ge \max_b \hat{Q}_b(t))\right)$$
$$\leq \frac{\epsilon}{K} \sum_{a \in \mathcal{A}} \Delta_a + \sum_{a \in \mathcal{A}} \Delta_a \lim_{n \to \infty} \frac{C}{n}$$
$$= \frac{\epsilon}{K} \sum_{a \in \mathcal{A}} \Delta_a.$$

As the upper and lower bound on the limit coincide, this proves the claim. It remains to show that $\sum_{t=1}^{\infty} \mathbb{P}(\hat{Q}_a(t) \ge \max_b \hat{Q}_b(t)) \le C < \infty$. Therefore, first note that

$$\begin{split} \mathbb{P}(\hat{Q}_a(t) \geq \max_b \hat{Q}_b(t)) &\leq \mathbb{P}(\hat{Q}_a(t) \geq \hat{Q}_{a*}(t)) \\ &\leq \mathbb{P}(\hat{Q}_a(t) \geq Q_a + \frac{\Delta_a}{2}) + \mathbb{P}(\hat{Q}_a(t) < Q_a + \frac{\Delta_a}{2}) \\ &= \mathbb{P}(\hat{Q}_a(t) \geq Q_a + \frac{\Delta_a}{2}) + \mathbb{P}(\hat{Q}_a(t) < Q_{a*} - \frac{\Delta_a}{2}) \\ &\leq 2 \max_a \mathbb{P}(|\hat{Q}_a(t) - Q_a| \geq \frac{\Delta_a}{2}) \end{split}$$

for the last equality note that $Q_a + \frac{\Delta_a}{2} = Q_{a^*} - \frac{\Delta_a}{2}$ by definition of $\Delta_a = Q_{a^*} - Q_{a^*}$ and in the second inequality we used that for two random variables X and Y it holds that

$$\mathbb{P}(X \ge Y) = \mathbb{P}(X \ge Y, Y \ge a) + \mathbb{P}(X \ge Y, Y < a) \le \mathbb{P}(X \ge a) + \mathbb{P}(Y < a).$$

For any arm a we will now prove that

$$\mathbb{P}(|\hat{Q}_a(t) - Q_a| \ge \frac{\Delta_a}{2}) \le \frac{\epsilon t}{K} \exp(-\frac{\epsilon t}{5K}) + \frac{16}{\Delta_a^2} \exp(-\frac{\Delta_a^2 \epsilon t}{16K}).$$

Then it is obvious that $\sum_{t=1}^{\infty} \mathbb{P}(\hat{Q}_a(t) \ge \max_b \hat{Q}_b(t)) \le C < \infty$. So, it holds

$$\begin{split} \mathbb{P}(|\hat{Q}_a(t) - Q_a| \geq \frac{\Delta_a}{2}) &= \sum_{s=1}^t \mathbb{P}(|\hat{Q}_a(t) - Q_a| \geq \frac{\Delta_a}{2}, T_a(t) = s) \\ &= \sum_{s=1}^t \mathbb{P}(|\hat{Q}_a(t) - Q_a| \geq \frac{\Delta_a}{2} | T_a(t) = s) \mathbb{P}(T_a(t) = s) \\ &\leq \sum_{s=1}^t 2 \exp(-\frac{\Delta_a^2 s}{8}) \mathbb{P}(T_a(t) = s), \end{split}$$

where we applied Hoeffdings inequality in the last step. We divide into two sums as follows.

Define $x = \lfloor \frac{\epsilon t}{2K} \rfloor$, then

$$\begin{split} \mathbb{P}(|\hat{Q}_{a}(t) - Q_{a}| \geq \frac{\Delta_{a}}{2}) &\leq \sum_{s=1}^{x} 2\exp(-\frac{\Delta_{a}^{2}s}{8}) \mathbb{P}(T_{a}(t) = s) + \sum_{s=x+1}^{t} 2\exp(-\frac{\Delta_{a}^{2}s}{8}) \mathbb{P}(T_{a}(t) = s) \\ &\leq \sum_{s=1}^{x} 2\mathbb{P}(T_{a}(t) = s) + \sum_{s=x+1}^{t} 2\exp(-\frac{\Delta_{a}^{2}s}{8}) \\ &\leq \sum_{s=1}^{x} 2\mathbb{P}(T_{a}(t) = s) + \frac{16}{\Delta_{a}^{2}}\exp(-\frac{\Delta_{a}^{2}x}{8}). \end{split}$$

In the last step we used that $\sum_{t=x+1}^{\infty} e^{-\kappa t} \leq \frac{1}{\kappa} e^{-\kappa x}$. Further, let $T_a^R(t)$ be the number of random explorations of the arm a before time t, then

$$\begin{split} \sum_{s=1}^{x} \mathbb{P}(T_{a}(t) = s) &\leq x \mathbb{P}(T_{a}^{R}(t) \leq x) \\ &\leq \frac{\epsilon t}{2K} \mathbb{P}(T_{a}^{R}(t) - \mathbb{E}[T_{a}^{R}(t)] \leq \lfloor \frac{\epsilon t}{2K} \rfloor - \frac{\epsilon t}{K}) \\ &\leq \frac{\epsilon t}{2K} \mathbb{P}(T_{a}^{R}(t) - \mathbb{E}[T_{a}^{R}(t)] \leq -\frac{\epsilon t}{2K}) \\ &= \frac{\epsilon t}{2K} \mathbb{P}(T_{a}^{R}(t) - \mathbb{E}[T_{a}^{R}(t)] \geq \frac{\epsilon t}{2K}) \\ &\leq \frac{\epsilon t}{2K} e^{-\frac{\epsilon t}{K10}}. \end{split}$$

The last inequality follows from Bernstein inequality and this is exactly what we wanted to prove. Bernstein inequality: Let X_i be i.i.d. r.v. with mean μ such that $|X_i - \mu| \leq M$ and $\mathbb{V}(\sum_{i=1}^n X_i) = \sigma^2$, then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}X_{i}-\mu\geq b\right)\leq\exp\left(\frac{\frac{1}{2}b^{2}}{\sigma^{2}+\frac{1}{3}Mb}\right).$$

In our case we have $b = \frac{\epsilon t}{2K}$, $\sigma^2 \leq \frac{t\epsilon}{K}$ and M = 1.