Prof. Dr. Leif Döring
Daniel Schmidt, Benedikt Wille **2. Exercise Sheet - Solutions**

Reinforcement Learning
02.03.2026

1. Upper bound on $\hat{Q}_a(t)$ for many samples

Suppose ν is a bandit model with 1-sub-gaussian arms. Show that under the UCB Algorithm $\hat{Q}_a(t) < Q_a + \Delta_a$ with probability $1-\delta$, given that $T_a(t) > \frac{2\log(1/\delta)}{\Delta_a^2}$.

Solution:

Proof: Consider w.l.o.g. that $\mathbb{P}^\pi\Big(T_a(t)=n\Big) > 0$ *for all* $n \in \{1,\dots,t-(k-1)\}$. *(UCB chooses each of the* $k-1$ *suboptimal arms at least once in the beginning). First we can observe that* $T_a(t) > \frac{2\log(1/\delta)}{\Delta_a^2}$ *is equivalent to* $\Delta_a > \sqrt{\frac{2\log(1/\delta)}{T_a(t)}}$*, so we will now consider the probability of* $\hat{Q}_a(t) - Q_a \geq \sqrt{\frac{2\log(1/\delta)}{T_a(t)}}$*. Then, considering the intersection with the condition* $T_a(t)=n$ *for some* $n \leq t-(k-1)$ *yields*

$$\begin{aligned}
&\mathbb{P}^\pi\Big(\hat{Q}_a(t) - Q_a \geq \sqrt{\frac{2\log(1/\delta)}{T_a(t)}} \cap (T_a(t)=n)\Big)\\
&= \mathbb{P}^\pi\Big(\frac{1}{T_a(t)}\sum_{i=1}^t X_i \mathbf{1}_{\{A_i=a\}} - Q_a \geq \sqrt{\frac{2\log(1/\delta)}{T_a(t)}} \cap (T_a(t)=n)\Big)\\
&= \mathbb{P}^\pi\Big(\frac{1}{n}\sum_{i=1}^t X_i \mathbf{1}_{\{A_i=a\}} - Q_a \geq \sqrt{\frac{2\log(1/\delta)}{n}} \cap (T_a(t)=n)\Big)\\
&= \mathbb{P}^\pi\Big(\frac{1}{n}\sum_{i=1}^t X_i \mathbf{1}_{\{A_i=a\}} - Q_a \geq \sqrt{\frac{2\log(1/\delta)}{n}} \Big| T_a(t)=n\Big)\mathbb{P}^\pi(T_a(t)=n)\\
&\leq \delta \mathbb{P}^\pi(T_a(t)=n).
\end{aligned}$$

Note that a conditional probability is still a probability measure so we can use the normal Hoeffdings inequality in the last step.

Furthermore, we obtain that

$$
\begin{aligned}
&\mathbb{P}^\pi\Big(\big(\hat{Q}_a(t) < Q_a + \Delta_a\big)\Big|\big(T_a(t) > \frac{2\log(1/\delta)}{\Delta_a^2}\big)\Big)\\
&\geq \mathbb{P}^\pi\Big(\hat{Q}_a(t) - Q_a < \sqrt{\frac{2\log(1/\delta)}{T_a(t)}}\Big|\big(T_a(t) > \frac{2\log(1/\delta)}{\Delta_a^2}\big)\Big)\\
&= \mathbb{P}^\pi\Big(\hat{Q}_a(t) - Q_a < \sqrt{\frac{2\log(1/\delta)}{T_a(t)}}\Big| \biguplus_{n=\lceil\frac{2\log(1/\delta)}{\Delta_a^2}\rceil}^{t-(k-1)} \big(T_a(t) = n\big)\Big)\\
&\geq \frac{\sum_{n=\lceil\frac{2\log(1/\delta)}{\Delta_a^2}\rceil}^{t-(k-1)} \mathbb{P}^\pi\Big(\hat{Q}_a(t) - Q_a < \sqrt{\frac{2\log(1/\delta)}{n}} \cap (T_a(t) = n)\Big)}{\sum_{n=\lceil\frac{2\log(1/\delta)}{\Delta_a^2}\rceil}^{t-(k-1)} \mathbb{P}^\pi\big(T_a(t) = n\big)}\\
&= \frac{\sum_{n=\lceil\frac{2\log(1/\delta)}{\Delta_a^2}\rceil}^{t-(k-1)} \mathbb{P}^\pi(T_a(t) = n) - \mathbb{P}^\pi\Big(\hat{Q}_a(t) - Q_a \geq \sqrt{\frac{2\log(1/\delta)}{n}} \cap (T_a(t) = n)\Big)}{\sum_{n=\lceil\frac{2\log(1/\delta)}{\Delta_a^2}\rceil}^{t-(k-1)} \mathbb{P}^\pi\big(T_a(t) = n\big)}\\
&\geq \frac{\sum_{n=\lceil\frac{2\log(1/\delta)}{\Delta_a^2}\rceil}^{t-(k-1)} \mathbb{P}^\pi(T_a(t) = n) - \delta\mathbb{P}^\pi(T_a(t) = n)}{\sum_{n=\lceil\frac{2\log(1/\delta)}{\Delta_a^2}\rceil}^{t-(k-1)} \mathbb{P}^\pi\big(T_a(t) = n\big)}\\
&= 1 - \delta,
\end{aligned}
$$

where we used the definition of conditional expectation and that $\mathbb{P}(A \cap B) = \mathbb{P}(B) - \mathbb{P}(A^c \cap B)$*.*

2. Regret bounds for UCB on σ-subgaussian bandit models

For σ-subgaussian bandit models the UCB exploration bonus is modified as

$$
\text{UCB}_a(t) := \begin{cases} \infty, & T_a(t) = 0, \\ \hat{Q}_a(t) + \sqrt{\frac{2\sigma^2\log(\frac{1}{\delta})}{T_a(t)}}, & T_a(t) \neq 0. \end{cases}
$$

Check that the regret bound in Theorem 1.3.8 using $\delta = \frac{1}{n^2}$ changes to

$$
R_n(\pi) \leq 3\sum_{a\in\mathcal{A}} \Delta_a + 16\sigma^2\log(n) \sum_{a:Q_a\neq Q_*} \frac{1}{\Delta_a},
$$

and that this leads to

$$
R_n(\pi) \leq 8\sigma\sqrt{Kn\log(n)} + 3\sum_{a\in\mathcal{A}} \Delta_A
$$

in Theorem 1.3.9.

Solution:

The general idea of the proof of 1.3.8 does not change. We start by defining G_m in the same way except that now

$$G_{2,m} := \left\{ \omega : \bar{Q}_m^{(a)}(\omega) + \sqrt{\frac{2\sigma^2 \log(\frac{1}{\delta})}{m}} < Q_{a_1} \right\}.$$

The proof that $G_m \subseteq H_m$ works exactly the same using the modified notion of G_m. Similarly, $\mathbb{P}(G_1^c) \leq n\delta$ still holds. We then choose $m = \left\lceil \frac{2\sigma^2 \log(1/\delta)}{1/4\Delta_a^2} \right\rceil$ in order to assure

$$\Delta_a - \sqrt{\frac{2\sigma^2 \log(\frac{1}{\delta})}{m}} \geq \frac{1}{2}\Delta_a.$$

By Hoeffdings inequality for σ-subgaussian random variables we obtain similarly to 1.3.8 that $\mathbb{P}(G_{2,m}^c) \leq \exp(-\frac{m\Delta_a^2}{8})$ and finally the regret bound in the exercise. As in 1.3.9, rewriting

$$R_n(\pi) \leq n\Delta + 3\sum_{a \in \mathcal{A}} \Delta_a + \frac{16K\sigma^2 \log(n)}{\Delta}$$

and optimizing over Δ leads to the alternative regret bound stated in the exercise.

3. **Best Baseline**

The variance of a random vector X is defined by to be $\mathbb{V}[X] := \mathbb{E}[||X||_2^2] - ||E[X]||_2^2$. Show by differentiation that

$$b_* = \frac{\mathbb{E}_{\pi_\theta}[X_A ||\nabla \log \pi_\theta(A)||_2^2]}{\mathbb{E}_{\pi_\theta}[||\nabla \log \pi_\theta(A)||_2^2]}$$

is the baseline that minimises the variance of the unbiased estimators

$$(X_A - b)\nabla \log(\pi_\theta(A)), \quad A \sim \pi_\theta,$$

of $\nabla J(\theta)$.

Solution:

We have

$$\begin{aligned}
&\mathbb{V}\Big((X_A - b)\nabla \log(\pi_\theta(A))\Big) \\
&= \mathbb{E}\Big[(X_A - b)^2 ||\nabla \log(\pi_\theta(A))||_2^2\Big] - \Big\|\mathbb{E}\Big[(X_A - b)\nabla \log(\pi_\theta(A))\Big]\Big\|_2^2 \\
&= \mathbb{E}\Big[(X_A - b)^2 ||\nabla \log(\pi_\theta(A))||_2^2\Big] - \Big\|\mathbb{E}\Big[X_A \nabla \log(\pi_\theta(A))\Big]\Big\|_2^2,
\end{aligned}$$

where we used the baseline trick in the last equation. We define $f(A) = ||\nabla \log(\pi_\theta(A))||_2$ to have a better overview. Then

$$\begin{aligned}
&\mathbb{V}\Big((X_A - b)\nabla \log(\pi_\theta(A))\Big) \\
&= \mathbb{E}\Big[(X_A - b)^2 f(A)^2\Big] - \Big\|\mathbb{E}\Big[X_A \nabla \log(\pi_\theta(A))\Big]\Big\|_2^2 \\
&= \mathbb{E}\Big[X_A^2 f(A)^2\Big] - 2b\mathbb{E}\Big[X_A f(A)^2\Big] + b^2 \mathbb{E}\Big[f(A)^2\Big] - \Big\|\mathbb{E}\Big[X_A \nabla \log(\pi_\theta(A))\Big]\Big\|_2^2.
\end{aligned}$$

We calculate the first derivative as

$$\frac{\partial \mathbb{V}\Big((X_A - b)\nabla \log(\pi_\theta(A))\Big)}{\partial b}$$
$$= -2\mathbb{E}\Big[X_A f(A)^2\Big] + 2b\mathbb{E}\Big[f(A)^2\Big].$$

Solving for the root gives

$$b* = \frac{\mathbb{E}\Big[X_A f(A)^2\Big]}{\mathbb{E}\Big[f(A)^2\Big]},$$

which is a minimum, as the second derivative $2\mathbb{E}\Big[f(A)^2\Big] \geq 0$ *almost surely. Plugging in the definition of* f *proves the claim.*