Prof. Dr. Leif Döring
Sara Klein, Benedikt Wille

Reinforcement Learning

3. Excercise Sheet

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* $n \in \{1, \dots, t-(k-1)\}$. *(We just go up to* $t-(k-1)$ *because we have to choose* $k-1$ *times a different arm as every arm has to be played once in the beginning.) First we obtain 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 first consider the probability of* $\hat{Q}_a(t) - Q_a \geq \sqrt{\frac{2\log(1/\delta)}{T_a(t)}}$. *Then, we first consider the intersection with condition* $T_a(t) = n$ *for some* $n \leq t-(k-1)$.

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

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

Further we obtain that

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

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. 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

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

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.*