Computer Engineering Department

Regret Bounds

Mohammad Hossein Rohban, Ph.D.

Spring 2025

Courtesy: Most of slides are adopted from https://annhe.xyz/2021/06/11/pac-rl/.

Def.

Here we define:

- Action-value: $Q(a) = \mathbb{E}[r|a]$
- Optimal value: $V^* = Q(a^*) = \max_a Q(a)$

How to define UCB?

For each arm a, we estimate an upper confidence bound $U_t(a)$, such that with high probability, $Q(a) \leq U_t(a)$. I.e., with high confidence, our estimate is an upper bound on the true arm value. The algorithm then, at each time step, selects the action with maximum UCB.

$\forall t, \forall a \quad Q(a) \leq U_t(a)$ w.h.p.

$\Leftrightarrow \quad Pr(\forall t \in [1,T] \; \forall a \quad Q(a) \leq U_t(a)) \geq 1-\delta$

At each time step, as long as all of the UCB bounds simultaneously hold, we're in good shape to prove sublinear regret. Why? Let's first do a proof-sketch so we can get a birds-eye view of how this will go.

Q(a*) upper-bounded by UCB

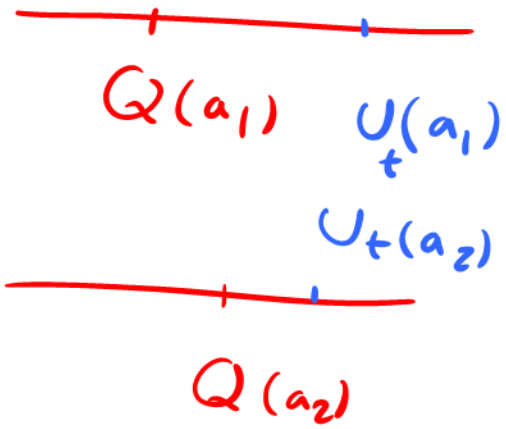
$a^* := \arg\max_a Q(a)$

If all UCB bounds hold, then no matter what action, a_t, we take, we have that $U_t(a_t) > Q(a^*)$. That is, the UCB of whatever action the algorithm takes is an upper bound on the optimal action. There are two cases.

$a_t := \arg\max_a U_t(a)$

Case 1: $a_t = a^*$

The action we select is actually the optimal action. Then $U_t(a_t) = U_t(a^*) > Q(a^*)$.



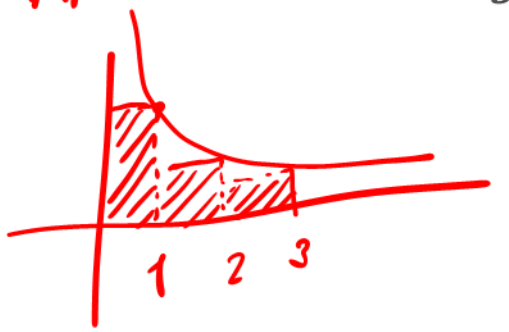
Case 2: $a_t \neq a^*$

Then $U_t(a_t) > U_t(a^*) > Q(a^*)$.

Hence Regret is upper-bounded

To see how this is relevant, let's look at the outline of the regret proof.

$$\sum_{n=1}^{N} \frac{1}{\sqrt{n}}$$



$$\begin{aligned} regret(UCB, T) &= \sum_{t=1}^{T} (Q(a^*) - Q(a_t)) \\ &= \sum_{t=1}^{T} U_t(a_t) - Q(a_t) + Q(a^*) - U_t(a_t) \\ &\leq \sum_{t=1}^{T} U_t(a_t) - Q(a_t) \end{aligned}$$

$U_t(a_t) \geq Q(a^*)$, so $Q(a^*) - U_t(a_t) \leq 0$.

$\rightarrow \hat{Q}(a_t) - Q(a_t) + d$

We have the inequality on the third line because the simultaneous UCB bounds give us $Q(a^*) - U_t(a_t) \leq 0$. Our estimate of $U_t(a_t)$ will be of the form $U_t(a_t) = \hat{Q}(a_t) + d$ where $\sum_{t=1}^{T} d_t$ will be a sublinear term. Ok, now onto the details.

How to bound $Q(a_i)$

Chernoff-Hoeffding Bound

Let $X_1, \ldots, X_n$ be i.i.d. random variables in $[0, 1]$ and let $\bar{X}_n = \frac{1}{n} \sum_{\tau=1}^{n} X_\tau$ be the sample mean. Then

$$P[E[X] > \bar{X}_n + u] \leq exp(-2nu^2)$$

In our interpretation, we let

$$P[Q(a_t) > \hat{Q}(a_t) + u] \leq exp(-2tu^2) = \frac{\delta}{t^2}$$

Where δ is a parameter and t is the current timestep and $n(a_t)$ is action selection count. We'll see later why setting the bound to be $\frac{\delta}{t^2}$ is useful.

Rewriting the Bound

$$P\left(\forall t \in [1,T]\ \forall a \quad Q(a_t) < \hat{Q}(a_t) + u\right)$$

$U_t(a_t)$

We use the Chernoff-Hoeffding equation to derive the design of the estimate $U(a_t)$.

Prob. that all Conf. bounds hold

Solving for u, we get,

$P(Q(a_t) > \hat{Q}(a_t) + u) \leq \frac{\delta}{t^2}$

$$u = \sqrt{\frac{1}{2n(a_t)} \log(t^2/\delta)}$$

So setting $U_t(a_t) = \hat{Q}(a_t) + \sqrt{\frac{1}{n(a_t)} \log(t^2/\delta)}$ gives us an UCB bound that holds with probability at least $1 - \frac{\delta}{t^2}$

$$P(B_1 \cup B_2 \cup \cdots \cup B_k) \leq \sum_{i=1}^{k} P(B_i)$$

All Bounds hold all the time

$Q(a) = \mathbb{E}(r|a)$

$\sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6} < 2$

The assumption we make for our sublinear regret bound to hold is that the Chernoff-Hoeffding Bounds hold for all arms at all time steps. Let's formally derive this quantity by looking at the probability of failure, i.e. the probability that at some timestep the bound for some arm is incorrect.

$$\bigcup_{t=1}^{T} Pr(Q(a^*) > U_t(a_t)) \leq \bigcup_{t=1}^{T} \bigcup_{i=1}^{m} Pr(|Q(a_t) - \hat{Q}(a_t)| > u)$$

$$\leq \sum_{t=1}^{T} \sum_{i=1}^{m} \frac{\delta}{t^2}$$

$$\leq 2m\delta$$

$P(\text{all bounds hold}) \geq 1 - 2m\delta$

All Bounds hold all the time

For the last inequality we used $\sum_{t=1}^{\infty} t^{-2} = \frac{\pi^2}{6} < 2$, which puts our analysis in the infinite horizon case.

Showing that the Chernoff bound gives us simultaneous success with probability at least $1 - 2m\delta$.

Deriving Regret

$$
\begin{aligned}
regret(UCB, T) &= \sum_{t=1}^{T} (Q(a^*) - Q(a_t)) \\
&= \sum_{t=1}^{T} U_t(a_t) - Q(a_t) + Q(a^*) - U_t(a_t) \\
&\leq \sum_{t=1}^{T} U_t(a_t) - Q(a_t)
\end{aligned}
$$

Deriving Regret

w.h.p at least $1-2m\delta$

$$\text{Regret} \leq \frac{1}{T}\sum_{t=1}^{T} U_t(a_t) - Q(a_t) = \sum_{t=1}^{T} \hat{Q}(a_t) + \sqrt{\frac{1}{2n(a_t)}\log(t^2/\delta)} - Q(a_t)$$

u : $U_t(a_t)$

$$\leq 2\sum_{t=1}^{T}\sqrt{\frac{1}{2n_t(a_t)}\log\frac{t^2}{\delta}}$$

$$\leq 2\sqrt{\frac{1}{2}\log\frac{T^2}{\delta}}\sum_{i=1}^{m}\sum_{n=1}^{n_t(i)}\sqrt{\frac{1}{n}}$$

$$\leq 2\sqrt{\frac{1}{2}\log\frac{T^2}{\delta}}\sum_{i=1}^{m}\sum_{n=1}^{T/m}\sqrt{\frac{1}{n}}$$

$$\leq 2\sqrt{\frac{1}{2}\log\frac{T^2}{\delta}}\sum_{i=1}^{m} 2\sqrt{T/m}$$

$$\leq 4\sqrt{\frac{1}{2}\log\frac{T^2}{\delta}Tm}$$

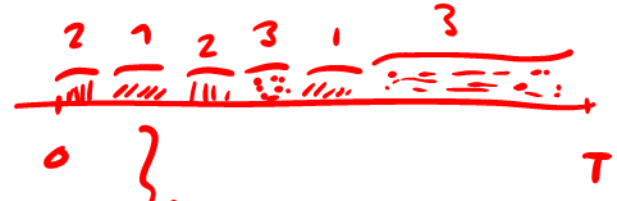
$\frac{1}{T}$

$\hat{Q}(a_t) - Q(a_t)$ U_t

$P(Q(a_t) < \hat{Q} + u) \geq 1 - \frac{\delta}{t^2}$

$P(\hat{Q}(a_t) < Q + u) \geq 1 - \frac{\delta}{t^2}$

$$\sum_{t=1}^{T}\sqrt{\frac{1}{n_t(a_t)}}$$



$$\sum_{j=1}^{m}\sum_{i=1}^{n(a_j)}\frac{1}{\sqrt{i}}$$

w.p. at least $1 - 2m\delta$.