Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Summary of Lecture 18: Regret Bounds

Summarized By: Benyamin Naderi

- **Regret** measures how much worse your agent performs compared to the best possible strategy (the optimal policy), over time. Formally, at each time step, there's a best possible reward the agent could have gotten (by acting optimally), and there's the reward it actually got. Regret is the sum of those differences over time.
- In learning procedure we prefer that this Regret to be sub-linear over time, which actually mean's that the aforementioned difference between optimal reward and the collected reward regarding to the agent's exploration, converges to zero in long-term.
- Let's remember some notations and , please be notified that in multi-armed bandit setting taking action means choosing the specified arm , and in this setting we are stateless so our value functions are just the function of the taken actions.
  - **Action-value:**
    $$Q(a) = \mathbb{E}[r \mid a]$$
    Expected reward when taking action a.
  - **Optimal value:**
    $$V^* = Q(a^*) = \max_a Q(a)$$
    The best possible expected reward across all actions.
- **How to define UCB?** As mentioned in the previous lecture, 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. in other words :
$$\begin{aligned} &\forall t, \forall a \quad Q(a) \leq U_t(a) \quad \text{w.h.p.} \\ \iff \quad &\Pr\left(\forall t \in [1, T], \forall a \quad Q(a) \leq U_t(a)\right) \geq 1 - \delta \end{aligned}$$
- If all UCB bounds hold, then for any action a_t taken, we have
$$U_t(a_t) > Q(a^*).$$
This means the UCB of the selected action upper bounds the optimal action's value.
There are two cases:
  - **Case 1:** $a_t = a^*$
    The selected action is the optimal action. Thus,
    $$U_t(a_t) = U_t(a^*) > Q(a^*).$$
  - **Case 2:** $a_t \neq a^*$
    Then,
    $$U_t(a_t) > U_t(a^*) > Q(a^*).$$

- now we try to derive an upper bound for regret which is not a function of $a^\star$, since basically we don't know it, To see how this is relevant, let's look at the outline of the regret proof:

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

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 $U_t(a_t)$ will be of the form

$$U_t(a_t) = \hat{Q}(a_t) + d_t$$

where $\sum_{t=1}^{T} d_t$ will be a sublinear term and the $\hat{Q}(a_t)$ is the empirical mean of reward of arms.

we can plug our estimate over $U_t(a_t)$ in the regret bound for furthur analysis.

Chernoff-Hoeffding Bound

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

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

In our interpretation, we let

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

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.
We use the Chernoff-Hoeffding equation to derive the design of the estimate $U(a_t)$. Solving for u, we get,

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

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 $1 - \frac{\delta}{t^2}$. or in other words using this u will gives us this bound :

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

Now we can rewrite the probability that all confidence bounds hold:

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

And remember that :

$$\mathsf{U}_t(a_t) = \hat{Q}(a_t) + u$$

Union Bound

For any countable collection of events $B_1, B_2, \ldots, B_k$,

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

- 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.

$$\begin{aligned} P\left(\bigcup_{t=1}^{T}\{Q(a^*) > U_t(a_t)\}\right) &\leq \sum_{t=1}^{T}\sum_{i=1}^{m} P\left(Q(a^\star) - \hat{Q}(a_t) > u\right) \\ &= \sum_{t=1}^{T}\sum_{i=1}^{m} \frac{\delta}{t^2} \\ &= m\delta \sum_{t=1}^{T} \frac{1}{t^2} \\ &\leq m\delta \cdot \frac{\pi^2}{6} \\ &\leq 2m\delta \end{aligned}$$

The probability that all confidence bounds hold simultaneously for all arms a_i and all time steps $t \in [1,T]$ is:

$$\begin{aligned} P\left(\bigcap_{t=1}^{T}\bigcap_{i=1}^{m}\left\{Q(a_i) \leq \hat{Q}_t(a_i) + u_t(a_i)\right\}\right) &= 1 - P\left(\bigcup_{t=1}^{T}\bigcup_{i=1}^{m}\left\{Q(a_i) > \hat{Q}_t(a_i) + u_t(a_i)\right\}\right) \\ &\geq 1 - \sum_{t=1}^{T}\sum_{i=1}^{m} P\left(Q(a_i) > \hat{Q}_t(a_i) + u_t(a_i)\right) \\ &= 1 - \sum_{t=1}^{T}\sum_{i=1}^{m} \frac{\delta}{t^2} \\ &\geq 1 - 2m\delta \end{aligned}$$

Given that we have simultaneous success of our bounds, let's now prove that the regret of the optimistic algorithm is indeed sublinear.

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

Now we plug in our estimator $U_t(a_t)$:

$$\begin{aligned}
\sum_{t=1}^{T} [U_t(a_t) - Q(a_t)] &= \sum_{t=1}^{T} \left[\hat{Q}_t(a_t) + \sqrt{\frac{1}{n_t(a_t)} \log\left(\frac{t^2}{\delta}\right)} - Q(a_t) \right] \\
&\leq 2 \sum_{t=1}^{T} \sqrt{\frac{1}{n_t(a_t)} \log\left(\frac{t^2}{\delta}\right)} \\
&= 2 \sqrt{\log\left(\frac{T^2}{\delta}\right)} \sum_{i=1}^{m} \sum_{n=1}^{n_T(i)} \frac{1}{\sqrt{n}} \\
&\leq 2 \sqrt{\log\left(\frac{T^2}{\delta}\right)} \sum_{i=1}^{m} 2\sqrt{T/m} \\
&= 4 \sqrt{mT \log\left(\frac{T^2}{\delta}\right)} \\
&= O\left(\sqrt{mT \log T}\right)
\end{aligned}$$

This bound holds with probability at least $1 - 2m\delta$.