Lecture #4: Stochastic bandits (Part 2)

Explore-then-Commit (input $N \in \mathbb{N}$)

Pull each arm N times.

Let $k^* \in \underset{k \in [K]}{\operatorname{argmax}} \ \hat{\mu}_k(NK)$

Play k^* until time T

Simple algorithm clearly separating exploration from exploitation.
Easy analysis.

Theorem ETC with $N = \frac{\ln T}{\Delta^2}$ satisfies

$$R_T \leq \sum_{\substack{k=1 \\ \Delta_k > 0}}^{K} \frac{\Delta_k}{\Delta^2} \ln T + \Delta_k.$$

Proof: Similarly to Greedy in the full info setting:

For $n = 1, \dots, N$ let $t_k(n)$ be the deterministic time where k is pulled for the n-th time.

By Hoeffding inequality:

$$\mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} X_k(t_k(n)) - X_{k^*}(t_k(n)) \geq 0\right) \leq e^{-N\Delta_k^2}$$

So:
$$R_T \leq \sum_{k=1}^{K} N\Delta_k + \sum_{k=1}^{K} \Delta_k (T-N) e^{-N\Delta_k^2}$$

Plugging the value of N:
$$R_T \leq \sum_{k=1}^{K} \frac{\Delta_k}{\Delta^2} \ln T + \Delta_k. \quad \square$$

Remark: • Again we can get a distribution-free bound scaling as $O((K \ln T)^{1/3} T^{2/3})$; and the instance dependent version requires knowledge of Δ

Two main drawbacks of these methods:

- they require knowledge of Δ.
- they scale in $\frac{1}{\Delta^2}$ ($\ln T^{2/3}$ in distribution-free bounds)

This is because they use a uniform exploration: each arm is explored the same amount of time.

exploration rounds depend on past observations.

A better strategy is to use an adaptive exploration: better arms are explored more often. The idea is that a very bad arm is quicker to detect as sub-optimal.

Successive Eliminations

Let $\mathcal{K} = [K]$

While $\mathrm{Card}(\mathcal{K}) > 1$:

Pull each arm in $\mathcal{K}$ once

For $k \in \mathcal{K}$:

if $\hat{\mu}_k(t) + \sqrt{\frac{2\ln T}{N_k(t)}} \leqslant \max_{k' \in \mathcal{K}} \hat{\mu}_{k'}(t) - \sqrt{\frac{2\ln T}{N_{k'}(t)}}$ then $\mathcal{K} \leftarrow \mathcal{K} \setminus \{k\}$

Pull the only arm in $\mathcal{K}$ until the end

Theorem: For SE, the regret satisfies for any $T \in \mathbb{N}$:

$$\boxed{R_T \leqslant \sum_{k, \Delta_k > 0} \left(\frac{32 \ln T}{\Delta_k} + 1 \right) + \frac{K}{T}}$$

Proof: Define the clean event

$$\mathcal{E} = \left\{ \begin{array}{l} \forall k \neq k^*, \forall t \in [T], \quad \hat{\mu}_k(t) - \mu_k \leqslant \sqrt{\frac{2\ln T}{N_k(t)}} \\ \qquad\qquad \forall t \in [T], \quad \hat{\mu}_{k^*}(t) - \mu_{k^*} \geqslant -\sqrt{\frac{2\ln T}{N_{k^*}(t)}} \end{array} \right\}$$

Thanks to our concentration lemma on $\hat{\mu}_k$:

$$P(\mathcal{E}) \geqslant 1 - K \sum_{t=1}^{T} \frac{t}{T^4} \geqslant 1 - \frac{K}{T^2}.$$

We now bound $\mathbb{E}[N_k(T)\mathbb{1}_{\{\mathcal{E}\}}]$.

Note that when $\mathcal{E}$ holds, we always have:

$$\hat{\mu}_{k^*}(t)+\sqrt{\frac{2\ln T}{N_{k^*}(t)}} \geq \mu_{k^*} \geq \mu_k \geq \hat{\mu}_k(t)-\sqrt{\frac{2\ln T}{N_k(t)}}.$$

So k^* is never eliminated from K.

For a suboptimal arm k, let N_k be the smallest integer such that:

$$4\sqrt{\frac{2\ln T}{N_k(t)}} \leq \Delta_k$$

i.e $\quad N_k = \left\lceil \frac{32\ln T}{\Delta_k^2} \right\rceil$.

Then once all arms in K have been pulled N_k times, we have if $\mathcal{E}$ holds

$$\hat{\mu}_k(t)+\sqrt{\frac{2\ln T}{N_k}} \leq \mu_k + 2\sqrt{\frac{\ln T}{N_k}} \leq \mu_{k^*} - 2\sqrt{\frac{\ln T}{N_k}} \leq \hat{\mu}_{k^*}(t) - \sqrt{\frac{\ln T}{N_k}}$$

So k is eliminated after at most N_k pulls if $\mathcal{E}$ holds:

$$\mathbb{E}[N_k(T)\mathbb{1}_{\mathcal{E}}] \leq \left\lceil \frac{32\ln T}{\Delta_k^2} \right\rceil$$

Finally: $$R_T \leq \sum_{k,\Delta_k>0} \Delta_k \left(\mathbb{E}\left[N_k(T) \mathbb{1}_{\mathcal{E}}\right] + \mathbb{E}\left[N_k(T)\, \mathbb{1}_{\text{not } \mathcal{E}}\right] \right)$$

$$\leq \sum_{k,\Delta_k>0} \Delta_k \sqrt{\frac{32 \ln T}{\Delta_k^2}} + T(1 - \mathbb{P}(\mathcal{E})))$$

$$\leq \sum_{k,\Delta_k>0} \left(32 \frac{\ln T}{\Delta_k} + 1\right) + \frac{K}{T}$$

Remarks • SE assumes a prior knowledge of T. assuming T is not too restrictive in practice, as we can use the doubling trick see exercise session #2

• we can easily get a better constant than 32

• This instance dependent bound also implies a distribution free bound $O(\sqrt{TK \ln T})$ see exercise session #2.

Upper Confidence Bound (UCB)

Pull each arm once

For $t \geq K+1$:

$$a_t \in \underset{k \in [K]}{\operatorname{argmax}} \underbrace{\hat{\mu}_k(t-1) + \sqrt{\frac{2 \ln(t)}{N_k(t-1)}}}_{\text{UCB score}}$$

- Greedy, but with UCB scores
  - $\rightarrow$ no underestimation of μ_k (with high probability)
- No prior knowledge of T.

Theorem

For any $T \in \mathbb{N}$, the regret of UCB satisfies

$$\boxed{R_T \leqslant \sum_{k, \Delta_k > 0} \left(8 \frac{\ln T}{\Delta_k} + 2 \right)}$$

Proof:

For $t \geqslant K+1$ and $k \neq k^*$, let

$$\mathcal{E}_{k,t} = \left\{ \begin{array}{l} \hat{\mu}_k(t) - \mu_k \leqslant \sqrt{\frac{2 \ln t}{N_k(t)}} \\ \hat{\mu}_{k^*}(t) - \mu_{k^*} \geqslant - \sqrt{\frac{2 \ln t}{N_{k^*}(t)}} \end{array} \right\}$$

$$\mathbb{P}(\mathcal{E}_t) \geqslant 1 - \frac{2}{t^3}$$

If $\mathcal{E}_{k,t}$ holds and $k \neq k^*$ is pulled at time t, then:

$$\hat{\mu}_k(t) + \sqrt{\frac{2 \ln t}{N_k(t-1)}} \geqslant \hat{\mu}_{k^*}(t) + \sqrt{\frac{2 \ln t}{N_{k^*}(t-1)}}$$

$\mathcal{E}_{k,t}$ holds, so $\quad \mu_k + 2\sqrt{\frac{2 \ln t}{N_k(t-1)}} \geqslant \hat{\mu}_k(t) + \sqrt{\frac{2 \ln t}{N_k(t-1)}}$

and $\quad \hat{\mu}_{k^*}(t) + \sqrt{\frac{2 \ln t}{N_{k^*}(t-1)}} \geqslant \mu_{k^*}$

In particular:

$$\mu_k + 2\sqrt{\frac{2\ln t}{N_k(t-1)}} \geq \mu_{k^*}$$

so $\left(\mathcal{E}_{k,t} \text{ and } a_t = k\right) \implies N_k(t-1) \leq \dfrac{8\ln t}{\Delta_k^2}$.

From here for $k \neq k^*$:

$$\mathbb{E}[N_k(T)] = 1 + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\left(a_t = k \text{ and } \mathcal{E}_{k,t}\right) + \mathbb{1}\left(a_t = k \text{ and not}(\mathcal{E}_{k,t})\right)\right]$$

$$\leq 1 + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\left(a_t = k \text{ and } N_k(t-1) \leq \frac{8\ln t}{\Delta_k^2}\right)\right] + 2\sum_{t=K+1}^{T} \frac{1}{t^3}.$$

$$\leq 1 + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{1}\left(a_t = k \text{ and } N_k(t-1) \leq \frac{8\ln T}{\Delta_k^2}\right)\right] + 2\int_1^{\infty} \frac{1}{s^3}\,ds$$

$$\leq 1 + \mathbb{E}\left[\left(\left\lfloor \frac{8\ln T}{\Delta_k^2} \right\rfloor + 1\right) - 1\right] + \left[-t^{-2}\right]_1^{\infty}$$

$$\leq 2 + \frac{8\ln T}{\Delta_k^2}. \qquad \boxtimes$$

- The $8\sum_{k, \Delta_k > 0} \dfrac{\ln T}{\Delta_k}$ instance dependent bound is nearly optimal. We'll see in exercise session that UCB can be made optimal with respect to the lower bound we are going to prove next week.

- UCB is said to use the optimism in the face of uncertainty principle: aiming at the best statistically possible scenario is a good strategy here.

- Previous algorithms/results hold for independent bounded rewards $X_k(t) \in [0,1]$

They can be easily extended to independent σ sub-Gaussian rewards, as similar concentration bounds hold.

eg UCB scores become

$$\hat{\mu}_k(t-1) + \sqrt{\frac{\sigma^2 \ln(t)}{2 N_k(t-1)}} \longrightarrow \text{same regret bounds, rescaled by } \sigma$$

What if σ is unknown? can be estimated see exercise session #3.