Lecture #4: Stochastic bandits (part 2)

Reminder: we proved $O\left(\frac{\ln T}{\Delta^2}\sum_a \Delta_a\right)$ bounds for ETC and ε-greedy

Remarks • the bound above is called instance dependent as it heavily relies on parameters of the instance Δ_k

A different choice of ε_t (or n) can lead to the following distribution-free bound for ε-Greedy:

$$R_T \leq O\left((K \ln T)^{1/3}\, T^{2/3}\right)$$

Two main drawbacks of ETC and ε-greedy

- they require knowledge of Δ.
- they scale in $\frac{1}{\Delta^2}$ (or $T^{2/3}$ 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 $\rightarrow$ adaptive version of ETC

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

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

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

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

if $\hat{\mu}_k(t) + \sqrt{\frac{2\ln T}{N_k(t)}} \leq \max_{k' \in \mathcal{K}} \hat{\mu}_{k'}(t) - \sqrt{\frac{2\ln T}{N_{k'}(t)}}$ 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 $T \in \mathbb{N}$:

$$\boxed{\mathbb{E}[R_T] \leq \sum_{k, \Delta_k > 0} \left(\frac{32 \ln T}{\Delta_k} + 1 \right) + \frac{K}{T}}$$

Proof: Define the clean event

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

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

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

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

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

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

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

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

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

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 lecture #4

• 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 end of lecture

• again this is a high probability bound

Upper Confidence Bound (UCB)

Pull each arm once

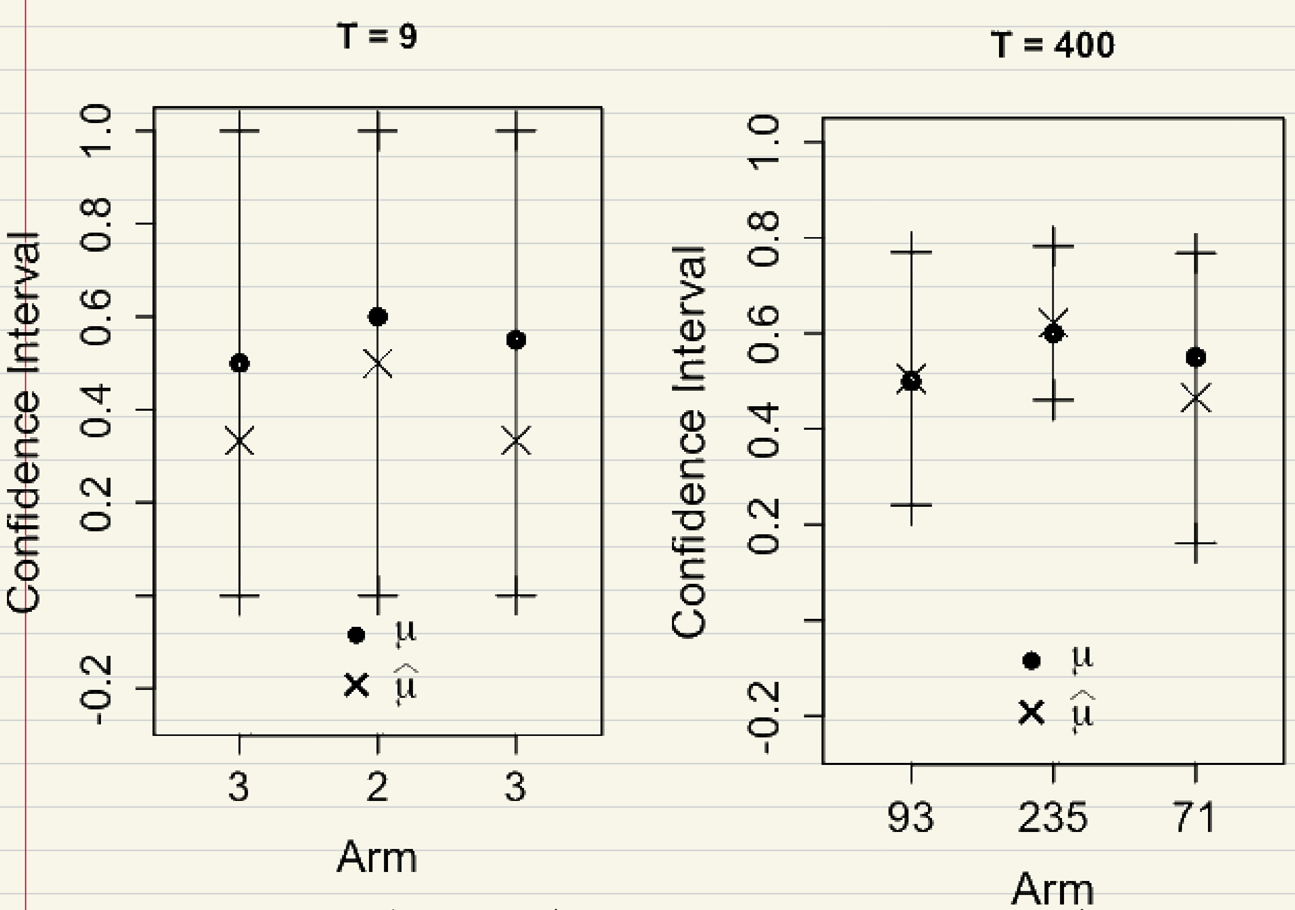
For $t \geq K+1$:

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

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

Idea of the algorithm:

- for each arm k, it builds a confidence interval on its expected reward based on past observation $I_k(t) = [L_k(t), U_k(t)]$.



- it is optimistic, acting as if the best possible rewards are real rewards.

- for rewards in $[0,1]$, we use a confidence upper bound

$$U_a(t) = \hat{\mu}_a(t-1) + \sqrt{\frac{2\ln t}{N_a(t-1)}}$$

Theorem

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

$$\mathbb{E}[R_T] \leq \sum_{k, \Delta_k > 0} \left(8 \frac{\ln T}{\Delta_k} + 2 \right)$$

Proof:

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

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

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

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

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

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

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

In particular:

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

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

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

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

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

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

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

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

- The $8\sum_{k,\Delta_k>0} \dfrac{\ln T}{\Delta_k}$ instance dependent bound is nearly optimal.

Modifications of UCB can be made to make it optimal

- 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?

✓ if σ unknown, but X_t bounded (with known bounds)

✓ if X_t is bounded $\in [m, M]$ with m, M unknown for $\sqrt{T}$ bound

✓ if X_t has a bounded Kurtosis: $\frac{\mathbb{E}[(X-\mathbb{E}[X])^4]}{\mathrm{Var}(X)^2} \leq K$ known

? general case

Until now, we only proved instance dependent bounds, i.e. bounds that depend on the bandits instance parameters Δ_a. But Δ_a can be very small, making these bounds explode. In such cases, we instead use distribution free bounds, which do not depend on any problem parameters (except T and K). They can actually be derived from the instance dep. bounds.

Distribution free bound. Let $\mathcal{B}$ be an arbitrary set of bandits. Suppose you are given a policy (algorithm) $\pi = \pi(T)$ designed for $\mathcal{B}$ that has the following guarantees

$$\mathbb{E}[N_k(T)] \leq C_0 + C\frac{\ln(T)}{\Delta_k^2}, \quad \forall \nu \in \mathcal{B}, \forall T \in \mathbb{N},$$

for some constants C_0, C.

1) First, show that it directly implies the following distribution free bound:

$$\mathbb{E}(R_T) \leq KC_0 + K\sqrt{CT\ln(T)}.$$

2) Show, with a refined analysis, that we even have the following bound

$$\mathbb{E}(R_T) \leq \sqrt{KT(C_0 + C\ln(T))}.$$

Solution: 1) Observe that $N_k(T) \leq T$, so that

$$\begin{aligned}\Delta_k\mathbb{E}[N_k(T)] &\leq C_0 + \min\left\{\Delta_k T, \frac{C\ln(T)}{\Delta_k}\right\} \\ &\leq C_0 + \sqrt{C\ln(T)T}.\end{aligned}$$

2) The finer analysis consists in saying that

$$\begin{aligned}\mathbb{E}(R_T) &= \sum_{k=1}^K \Delta_k\mathbb{E}[N_k(T)] \\ &\leq \sum_{k=1}^K \min\left\{\Delta_k\mathbb{E}[N_k(T)], C_0 + \frac{C\ln(T)}{\Delta_k}\right\} \\ &\leq \sum_{k=1}^K \sqrt{\mathbb{E}[N_k(T)]}\sqrt{C_0 + C\ln(T)} \\ &\leq \sqrt{C_0 + C\ln(T)}\sqrt{K\sum_{k=1}^K \mathbb{E}[N_k(T)]} \qquad \text{Cauchy Schwarz} \\ &\leq \sqrt{KT(C_0 + C\ln(T))}.\end{aligned}$$

Also, some algorithms assume knowledge of T. Not a big deal, because otherwise, we can still use the doubling trick.

Doubling trick. This exercise analyses a meta-algorithm based on the doubling trick that converts a policy depending on the horizon to a policy with similar guarantees that does not. Let $\mathcal{B}$ be an arbitrary set of bandits. Suppose you are given a policy (algorithm) $\pi = \pi(T)$ designed for $\mathcal{B}$ that accepts the horizon T as a parameter and has a regret guarantee of

$$\max_{1 \le t \le T} \mathbb{E}\left(R_t(\pi(n), \nu)\right) \le f_T(\nu), \quad \forall \nu \in \mathcal{B}.$$

For a fixed sequence of integers $T_1 < T_2 > T_3 < \ldots$, we define the algorithm $\tilde{\pi}$ that first runs $\pi(T_1)$ on $[\![1, T_1]\!]$; then runs **independently** $\pi(T_2)$ on $[\![T_1, T_1 + T_2]\!]$; etc. So $\tilde{\pi}$ runs $\pi(T_i)$ on $[\![\sum_{j=1}^{i-1} T_j, \sum_{j=1}^{i} T_j]\!]$ and does not require a prior knowledge of T.

1) For a fixed $T \in \mathbb{N}$, let $\ell_{\max} = \min\{\ell \in \mathbb{N}^* \mid \sum_{i=1}^{\ell} T_i \ge T\}$. Prove that for any $\nu \in \mathcal{B}$, the regret of $\tilde{\pi}$ on ν is at most

$$\mathbb{E}\left(R_T(\tilde{\pi}, \nu)\right) \le \sum_{\ell=1}^{\ell_{\max}} f_{T_\ell}(\nu).$$

2) (Distribution free bound) Suppose that $f_T(\nu) \le \sqrt{T}$. Show that for a good choice of n_ℓ, for any $\nu \in \mathcal{B}$ and $T \in \mathbb{N}$:

$$\mathbb{E}\left(R_T(\tilde{\pi}, \nu)\right) \le \frac{1}{\sqrt{2} - 1} \sqrt{T}.$$

3) (Instance dependent bound) Suppose that $f_T(\nu) \le g(\nu) \ln(T)$ for some function g. Show that with the same choice of sequence n_ℓ as in b), we can bound the regret for any $\nu \in \mathcal{B}$ and $T \in \mathbb{N}$ as:

$$\mathbb{E}\left(R_T(\tilde{\pi}, \nu)\right) \le g(\nu) \frac{\ln(T)^2}{2\ln(2)}.$$

4) Can you suggest a sequence of n_ℓ such that for some universal constant $C > 0$, the regret of $\tilde{\pi}$ can be bounded for any $\nu \in \mathcal{B}$ and $T \in \mathbb{N}$ as:

$$\mathbb{E}\left(R_T(\tilde{\pi}, \nu)\right) \le C g(\nu) \ln(T).$$

Solution: 1) is by definition of $\tilde{\pi}$.
2) is for the choice $T_\ell = 2^\ell$.
3) directly derives from the choice of n_ℓ.
4) $T_\ell = 2^{2^\ell}$.