Lecture #3: Stochastic bandits

Basic algorithms

General Setting / online learning

At each round $t = 1, \dots, T$:

- agent observes a context $c_t \in \mathcal{X}$ (optional step)
- agent chooses an action $a_t \in K$ (possibly at random)
- environment chooses a loss function $f_t : K \to \mathbb{R}_+$
- agent suffers loss $f_t(a_t)$ and observes
  - the losses of every action $f_t(x)$ $\forall x \in K$ → full information feedback
  - the loss of the chosen action only: $f_t(a_t)$ → bandit feedback

The goal of the player is to minimise his cumulative loss: $\hat{L}_T = \sum_{t=1}^{T} f_t(a_t)$

Stochastic bandits setting (random table model)

At each round $t = 1, \dots, T$:

- agent picks an arm $a_t \in \{1, \dots, K\}$ (possibly at random)
- observes and gets reward $X_{a_t}(t) \overset{iid}{\sim} \nu_{a_t}$

goal: maximise cumulative reward

→ only observe the reward of the pulled arm

→ $\underbrace{\text{exploration}}_{\text{estimate optimal arm by pulling all arms}}$ vs $\underbrace{\text{exploitation}}_{\text{maximize reward by pulling arm which seems the best}}$ trade-off

→ rewards are iid. $X_k(t) \sim \nu_k$ with $\mathbb{E}[X_k(t)] = \mu_k$.

Regret definition?

$$\hat{R}_T = \max_{k \in [K]} \sum_{t=1}^{T} X_k(t) - \sum_{t=1}^{T} X_{a_t}(t) \quad ?$$

Consider K arms with $\nu_k = \text{Bernoulli}\left(\frac{1}{2}\right)$.

→ All arms are the same, there is no bad choice and no bad algorithm

but
$$\mathbb{E}\left[\hat{R}_T\right] = \mathbb{E}\left[\max_{k \in [K]} \sum_{t=1}^{T} \left(X_k(t) - \frac{1}{2}\right)\right]$$

$$\gtrsim \sqrt{T \ln K}$$ | see lower bound of online learning with experts.

We want a regret notion that does not blow up with stochastic fluctuations

(Pseudo)-regret definition

$$R_T = \max_{k \in [K]} \sum_{t=1}^{T} \mu_k - \sum_{t=1}^{T} \mu_{a_t}$$ (still an r.v.)

Notations:

- $\mu^* = \max_k \mu_k$
- $\Delta_k = \mu^* - \mu_k$ $\begin{cases} > 0 & \text{for sub-optimal arms} \\ = 0 & \text{for optimal arms} \end{cases}$
- $\Delta = \min_{k, \Delta_k > 0} \Delta_k$
- $N_k(t) = \sum_{s=1}^{t} \mathbb{1}_{\{a_s = k\}}$ number of pulls on arm k.

Lemma: (Regret decomposition)

For any policy, $R_T = \sum_{k=1}^{K} \Delta_k \quad N_k(T)$

Proof:

$$
\begin{aligned}
R_T &= \sum_{t=1}^{T} \mu^* - \mu_{a_t} \\
&= \sum_{t=1}^{T} \sum_{k=1}^{K} (\mu^* - \mu_k) \mathbb{1}_{a_t = k} \\
&= \sum_{k=1}^{K} \Delta_k \sum_{t=1}^{T} \mathbb{1}_{a_t = k} \\
&= \sum_{k=1}^{K} \Delta_k \quad N_k(T)
\end{aligned}
$$

$\square$.

Bounding the regret $\iff$ Bounding number of pulls of bad arms

Random table model vs. stack of rewards

Observe and get reward	observe and get reward
$X_{a_t}(t)$	$X_{a_t}(N_{a_t}(t))$

We can show that both models are equivalent

we will sometimes use the stack of rewards model in the proof (easier to analyse)

Variants and extensions

- Contextual bandit: $X_k(t) \sim \nu_k(c_{k,t})$ for a known context c_t.
- Linear bandit: $\nu_k(c_t) = \mathcal{N}(x^T c_{k,t}, \sigma^2)$
- Structured bandits: the algorithm knows constraints on $(\mu_k)_{k \in K}$ (e.g. Lipschitz, linear, monotone, ...)

Other objectives

pure exploration problems

see lecture 6

- Minimise simple regret Δ_{a_T}
- Best arm identification: return an arm at time T and maximise the probability it is a best arm

Algorithmic idea estimate the arm means with the empirical means.

Question If I have 10 samples on arm k, how reliable is my estimate for μ_k?

Additional notation

- $$\hat{\mu}_k(t) = \frac{1}{N_k(t)} \sum_{s=1}^{t} X_k(s)\, \mathbb{1}_{\{a_s = k\}}$$ (empirical mean).

Follow the leader algorithm

For $t = 1, \dots, K$:

$\quad a_t = t$

For $t \geq K+1$:

$$a_t \in \underset{k \in [K]}{\text{argmax}} \; \hat{\mu}_k(t-1)$$

Theorem For $\nu_1 = \text{Ber}\left(\frac{3}{4}\right)$, $\nu_2 = \text{Ber}\left(\frac{1}{4}\right)$, Greedy satisfies in the bandit setting:

$$R_T \geq \frac{T-1}{32}.$$

Proof:

$$\mathbb{P}(X_1(1) = 0, X_2(2) = 1) = \left(\frac{1}{4}\right)^2 = \frac{1}{16}$$

If $X_1(1) = 0$ and $X_2(2) = 1$, Greedy will keep pulling the arm 2 until T, so that:

$$\mathbb{E}[N_2(T)] \geq \frac{T-1}{16}. \qquad \square$$

Greedy does not explore enough.

It can underestimate the optimal arm and never pull it again

Law of large numbers and central limit theorem are not strong enough tool for controlling the accuracy of the estimates $\hat{\mu}_a(t)$ at finite times. (They are asymptotic results)

→ use of concentration inequalities.

Hoeffding inequality

Let $(X_t)_{t \geq 1}$ be a sequence of independent random variables that are a.s in $[a, b]$,

then for all $\varepsilon > 0$

$$\mathbb{P}\left(\sum_{s=1}^{t} X_s - \mathbb{E}\left[\sum_{s=1}^{t} X_s\right] \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{t(b-a)^2}\right).$$

Equivalently for all $\delta \in (0,1)$

$$\mathbb{P}\left(\sum_{s=1}^{t} X_s - \mathbb{E}\left[\sum_{s=1}^{t} X_s\right] \geq (b-a)\sqrt{\frac{t}{2}\ln\left(\frac{1}{\delta}\right)}\right) \leq \delta.$$

Proof for sub-Gaussian random variables

Definition

a r.v. X_s is σ^2-sub-Gaussian if for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda(X_s - \mathbb{E}[X_s])}\right] \leq e^{\frac{1}{2}\sigma^2\lambda^2}$$

Exercise: bounded in $[a,b]$ implies $\frac{(b-a)^2}{4}$-sub-Gaussian

Proof: (of Hoeffding inequality)

Let $S_t = \sum_{s=1}^{t} X_s - \mathbb{E}[X_s]$

by independence, for any $\lambda \in \mathbb{R}$

sub-Gaussian

$$\mathbb{E}\left[e^{\lambda S_t}\right] = \prod_{s=1}^{t} \mathbb{E}\left[e^{\lambda(X_s - \mathbb{E}[X_s])}\right] \leq e^{\frac{t(b-a)^2\lambda^2}{8}}$$

$$\mathbb{P}(S_t \geq \varepsilon) = \mathbb{P}\left(e^{\lambda S_t} \geq e^{\lambda\varepsilon}\right)$$

$$\leq \mathbb{E}\left[e^{\lambda S_t}\right] e^{-\lambda\varepsilon}$$

Markov inequality

$$\leq e^{\frac{t(b-a)^2\lambda^2}{8} - \lambda\varepsilon}$$

Minimizing over $\lambda \in \mathbb{R}$, $\lambda^* = \frac{4\varepsilon}{t(b-a)^2}$

$$P(S_t \geq \varepsilon) \leq e^{-\frac{2\varepsilon^2}{t(b-a)^2}}$$

$\square$

⚠ Warning

In the analysis of algorithms, we want concentration bounds on $\hat{\mu}_k(t) - \mu_k = \frac{1}{N_k(t)} \sum_{s=1}^{N_k(t)} X_k(s)$

stack of rewards model.

Hoeffding inequality does not apply directly!

- $N_k(t)$ is a random variable, that is **not** independent from the $X_k(s)$, $s < N_k(t)$

Explore-then-Commit algorithm

parameter $n \in \mathbb{N}^*$

For $t = 1, \dots, nK$: explore by drawing each arm n times.

For $t \geq nK+1$:

pull the best empirical arm until the end, i.e.

$$a_t = \operatorname{argmax}_k \hat{\mu}_k(nK)$$

Simple algorithm clearly separating exploration from exploitation. Easy analysis.

Theorem:

If all distributions ν_k are bounded in $[0,1]$ and $1 \leq n \leq T/K$, then ETC has expected regret

$$\mathbb{E}[R_T] \leq n \sum_{k=1}^{K} \Delta_k + (T - nK) \sum_{k=1}^{K} \Delta_k \exp(-n\Delta_k^2)$$

Proof:

$$R_T = \sum_{k=1}^{K} \Delta_k N_k(T)$$

if $n \leq T/K$,
$$N_k(T) = \begin{cases} n & \text{if } k \neq \operatorname{argmax}_{k'} \hat{\mu}_{k'}(nK) \\ n + (T - nK) & \text{if } k = \operatorname{argmax}_{k'} \hat{\mu}_{k'}(nK) \end{cases}$$

$$R_T \leq n\sum_{k=1}^{K} \Delta_k + (T-nK)\sum_{k=1}^{K} \Delta_k \, \mathbb{1}_{k=\arg\max_{k'} \hat{\mu}_{k'}(nK)}.$$

Let $k^* = \arg\max_k \mu_k$ $(\mu_{k^*} = \mu^*)$

$$\mathbb{E}[R_T] \leq n\sum_{k=1}^{K} \Delta_k + (T-nK)\sum_{k=1}^{K} \Delta_k \, \mathbb{P}\left(k = \arg\max_{k'} \hat{\mu}_{k'}(nK)\right)$$

$$\leq n\sum_{k=1}^{K} \Delta_k + (T-nK)\sum_{k=1}^{K} \Delta_k \, \mathbb{P}\left(\hat{\mu}_k(nK) \geq \hat{\mu}_{k^*}(nK)\right)$$

$$\mathbb{P}\left(\hat{\mu}_k(nK) \geq \hat{\mu}_{k^*}(nK)\right) = \mathbb{P}\left(\sum_{s=1}^{n} X_k(s) - \sum_{s=1}^{n} X_{k^*}(s) \geq 0\right)$$

$$= \mathbb{P}\left(\sum_{s=1}^{n}(X_k(s) - \mu_k) - \sum_{s=1}^{n}(X_{k^*}(s) - \mu^*) \geq n\Delta_k\right).$$

Hoeffding: $\leq e^{-n\Delta_k^2}$ $\square$.

- n too large $\rightarrow$ explore too much
- n too small $\rightarrow$ not enough exploration, might pull suboptimal arm for $T-nK$ steps.

what n should we choose?

for $\Delta = \min_{k, \Delta_k > 0} \Delta_k$ and $n = \left\lceil \frac{\ln(T)}{\Delta^2} \right\rceil$

$$\boxed{\mathbb{E}[R_T] \leq \sum_{k=1}^{K} \frac{\Delta_k \ln T}{\Delta^2} + \sum_{k=1}^{K} \Delta_k.}$$

ε-Greedy

sequence of probabilities ε_t.

For $t = 1, \dots, K$:

$$a_t = t$$

For $t \geq K+1$:

$$\begin{cases} \text{with proba } \varepsilon_t, & a_t \sim \mathcal{U}([K]) \\ \text{with proba } 1-\varepsilon_t, & a_t \in \underset{k \in [K]}{\operatorname{argmax}} \hat{\mu}_k(t-1) \end{cases}$$

explore uniformly at random

- used in practice because of its simplicity.
- FTL with forced exploration $\rightarrow$ don't get stuck underestimating μ_{k^*}

We can show that setting $\varepsilon_t \approx \frac{K}{\Delta^2 t}$, it gets an expected regret bound

$$\mathbb{E}[R_T] = O\left(\sum_{k=1}^{K} \frac{\Delta_k}{\Delta^2} \ln T \right)$$

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

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

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

• the instance dependent bound requires a prior knowledge of Δ, which is usually unknown.

→ these 2 remarks are also valid for ETC

Two main drawbacks of these methods:

- 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.

Upper Confidence Bound (UCB)

Pull each arm once

For $t \geq K+1$:

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

- FTL, but with UCB scores

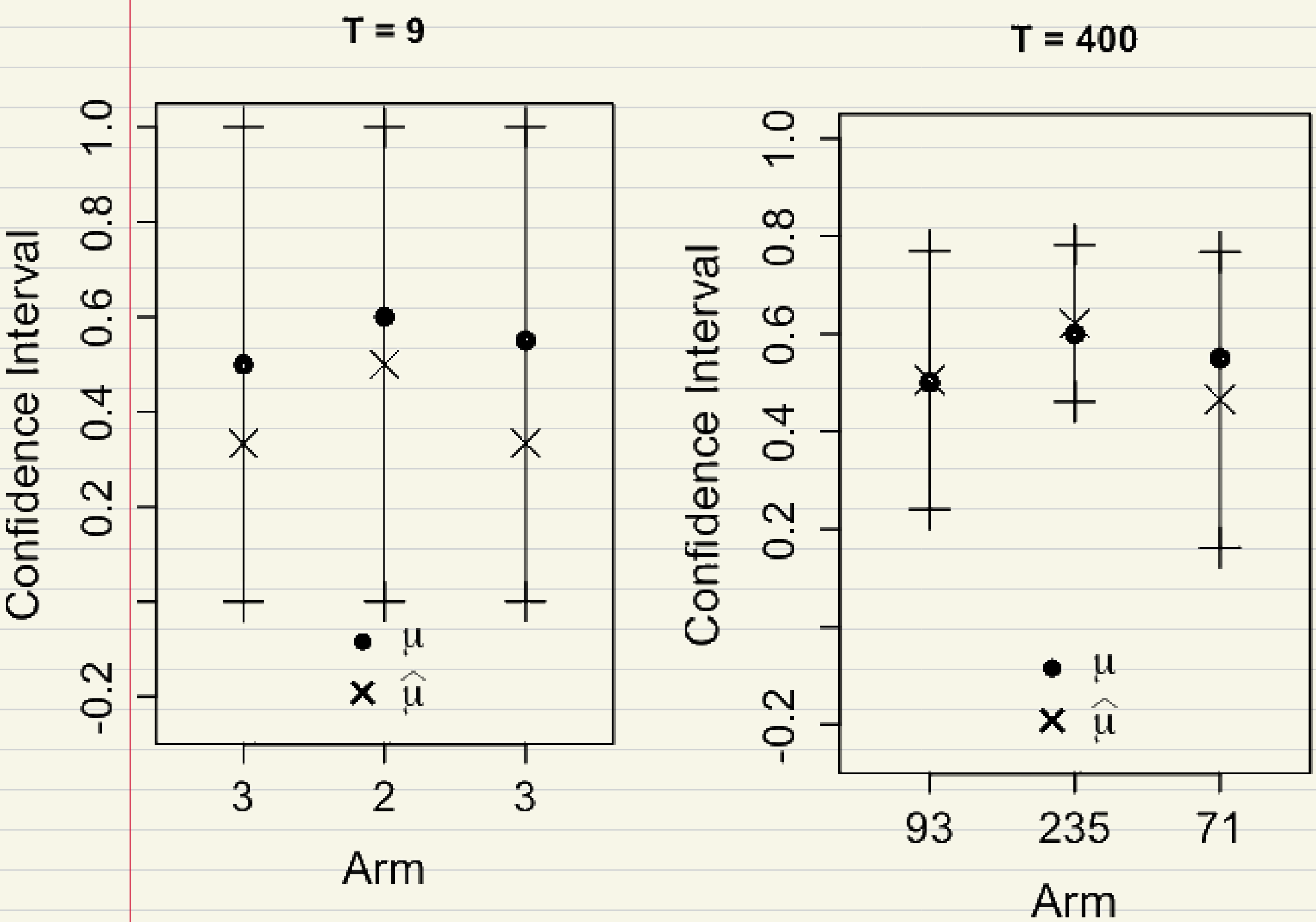
  → no underestimation of μ_k (with high probability)

- No prior knowledge of T (nor Δ)

- Optimism in the face of uncertainty

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

$$U_k(t) = \hat{\mu}_k(t-1) + \sqrt{\frac{2\ln t}{N_k(t-1)}}$$

because of the following concentration inequality:

Lemma: (bandit concentration)

For any bandit algorithm, any $k \in [K]$, $t \in \mathbb{N}$, $\delta \in (0,1)$ and distributions ν_a in $[0,1]$:

$$\mathbb{P}\left(\mu_k - \hat{\mu}_k(t) \geq \sqrt{\frac{\ln(1/\delta)}{2N_k(t)}}\right) \leq t\delta.$$

$$\mathbb{P}\left(\hat{\mu}_k(t) - \mu_k \geq \sqrt{\frac{\ln(1/\delta)}{2N_k(t)}}\right) \leq t\delta.$$

Recall this is not a trivial consequence of Hoeffding inequality, $N_k(t)$ is a random variable and $\hat{\mu}_k(t), N_k(t)$ are not independent!

Proof (for stack of rewards model):

$$\mathbb{P}\left(\hat{\mu}_k(t) - \mu_k \geq \sqrt{\frac{\ln(1/\delta)}{2N_k(t)}}\right) = \sum_{n=1}^{t} \mathbb{P}\left(\hat{\mu}_k(t) - \mu_k \geq \sqrt{\frac{\ln(1/\delta)}{2n}} \text{ and } N_k(t) = n\right)$$

$$\overset{\text{stack of rewards}}{=} \sum_{n=1}^{t} \mathbb{P}\left(\sum_{s=1}^{n}(X_k(s) - \mu_k) \geq \sqrt{\frac{n\ln(1/\delta)}{2}} \text{ and } N_k(t) = n\right)$$

$$\leq \sum_{n=1}^{t} \mathbb{P}\left(\sum_{s=1}^{n}(X_k(s) - \mu_k) \geq \sqrt{\frac{n\ln(1/\delta)}{2}}\right)$$

$$\overset{\text{Hoeffding}}{\leq} t\delta. \quad \square$$

For UCB, we thus have $U_k(t) \geq \mu_k$ with probability larger than $1 - \frac{1}{t^3}$.

"$\hat{\mu}_k(t-1) + \sqrt{\frac{2\ln t}{N_k(t-1)}}$"

Theorem (Regret UCB)

If the distributions ν_k have supports in $[0,1]$, then for UCB and all k s.t. $\Delta_k < 0$:

$$\mathbb{E}[N_k(T)] \leqslant \frac{8 \ln T}{\Delta_k^2} + 2.$$

In particular, it implies the regret bound for UCB

$$\mathbb{E}[R_T] \leqslant \sum_{k, \Delta_k > 0} \frac{8 \ln T}{\Delta_k} + 2\Delta_k.$$

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:

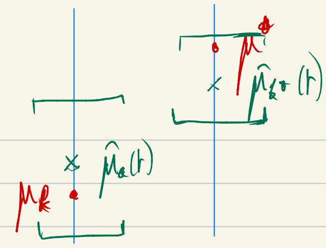
$$\hat{\mu}_k(t) + \sqrt{\frac{2 \ln t}{N_k(t-1)}} \geqslant \hat{\mu}_{k^*}(t) + \sqrt{\frac{2 \ln t}{N_{k^*}(t-1)}}$$

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

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

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

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

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

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

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

- The $8\sum_{k, \Delta_k > 0} \dfrac{\ln T}{\Delta_k}$ instance dependent bound is nearly optimal (see lecture 6)

- 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

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

- all the presented bounds are instance dependent bounds (i.e. they depend on Δ_k), and become insignificant for $\Delta_k \to 0$

Theorem (UCB distribution-free bound)

If the distributions ν_k have supports in $[0,1]$, then the regret of UCB is bounded as:

$$\mathbb{E}[R_T] \leq K\sqrt{8 T \ln T} + 2K$$

→ distribution free

→ can be improved to (nearly optimal) bound $\mathcal{O}(\sqrt{KT\ln T})$ (left as exercise).

Proof:

$$\mathbb{E}[N_k(T)] \leq \min\left(\frac{8\ln T}{\Delta_k^2} + 2, T\right) \leq \min\left(\frac{8\ln T}{\Delta_k^2}, T\right) + 2$$

so that

$$\Delta_k \, \mathbb{E}[N_k(T)] \leq \min_{\Delta > 0} \left(\frac{8 \ln T}{\Delta}, T\Delta \right) + 2$$

$$= \sqrt{8 T \ln T} + 2$$

$$\mathbb{E}[R_T] = \sum_k \Delta_k \, \mathbb{E}[N_k(T)] \leq K \sqrt{8 T \ln T} + 2K$$

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 $\hat{\mu}_k(t) + \sqrt{\frac{2 \ln T}{N_k(t)}} < \max_{k' \in \mathcal{K}} \hat{\mu}_{k'}(t) - \sqrt{\frac{2 \ln T}{N_{k'}(t)}}$ 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, if the distributions ν_k are bounded in $[0,1]$

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

Proof: Define the clean event

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

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

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

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

Remarks

- SE assumes a prior knowledge of T. assuming T is not too restrictive in practice, as we can use the doubling trick

- can be useful for some applications, as exploration and exploitation are clearly separated.