Lecture #3: Stochastic bandits (part 1)

Full Information Setting

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

- agent picks an arm $a_t \in \{1, \dots, K\}$ (possibly at random)
- observes reward vector $X(t) \in [0,1]^K$

  gets reward $X_{a_t}(t)$.

$$R_T = \max_{k \in [K]} \sum_{t=1}^{T} X_k(t) - \sum_{t=1}^{T} X_{a_t}(t)$$

a_t is $\sigma(U, X(1), \dots, X(t-1))$ measurable

U: possible randomisation

As learning with experts, but:
- rewards instead of loss ($\ell_t \longleftrightarrow 1 - X(t)$)
- choose pure actions (K-simplex $\longleftrightarrow \{1, \dots, K\}$)
  but can randomise over actions.

The $X(t)$ were chosen adversarially (worst case) in 1st lecture.

What if instead they are stochastic?

Assume
- $(X(t))_t$ are i.i.d.
- $X_k(t) \sim \nu_k$ with $\mathbb{E}[X_k(t)] = \mu_k$.

Problem should be easier?

$\hookrightarrow$ not really: we proved the lower bound in this setting:

for any algorithm, with $X_k(t) \sim \mathrm{Ber}\left(\frac{1}{2}\right)$

$$\mathbb{E}[R_T] \gtrsim \sqrt{\frac{T}{2} \ln K}$$

However, we can have much better results with the **pseudo-regret**:

$$\bar{R}_T = \max_{k \in [K]} \sum_{t=1}^{T} \mu_k - \sum_{t=1}^{T} \mu_{a_t}$$

↳ expectation wrt the realizations of $X(t)$,
but still a random variable!

Previous example yields $\bar{R}_T = 0$. Makes sense: we cannot guess in advance heads or tails.

⚠ Warning: $\mathbb{E}[R_T] \neq \mathbb{E}[\bar{R}_T]$

Actually, $\mathbb{E}[R_T] > \mathbb{E}[\bar{R}_T]$. Why?

$$\boxed{\bar{R}_T = T \max_k \mu_k - \sum_{t=1}^{T} \mu_{a_t}}$$

→ from now on, we will write R_T for the pseudo-regret.

Notations:

- $\mu^* = \max_k \mu_k$
- $\Delta_k = \mu^* - \mu_k \begin{cases} > 0 & \text{for sub-optimal arms} \\ = 0 & \text{for optimal arms} \end{cases}$
- $\Delta = \min_{k, \Delta_k > 0} \Delta_k$

$\cdot N_k(t) = \sum_{s=1}^{t} \mathbb{1}_{\{a_t = k\}}$ number of pulls on arm k.

Lemma:

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

Proof:

$$R_T = \sum_{t=1}^{T} \mu^* - \mu_{a_t}$$

$$= \sum_{t=1}^{T} \mu^* - \sum_{k=1}^{K} \mathbb{1}_{a_t=k} \mu_k$$

$$= \sum_{k=1}^{K} \sum_{t=1}^{T} (\mu^* - \mu_k) \quad \mathbb{1}_{a_t=k}$$

$$= \sum_{k=1}^{K} \Delta_k \quad \sum_{t=1}^{T} \mathbb{1}_{a_t=k}$$

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

□.

Greedy algorithm (or Follow The Leader)

Choose a_1 arbitrarily

For $t \geqslant 2$:

$$a_t \in \operatorname{argmax}_{k \in [K]} \sum_{s=1}^{t-1} X_k(s)$$

Theorem For any $(\mu_1, \dots, \mu_K) \in [0,1]^K$ and $T \in \mathbb{N}$, Greedy satisfies in the Full Information setting:

$$\mathbb{E}[R_T] \leqslant \sum_{k, \Delta_k > 0} \frac{1}{\Delta_k}$$

Proof: $\mathbb{E}[R_T] = \sum_{k=1}^{K} \Delta_k \, \mathbb{E}[N_k(T)]$

Let us bound $\mathbb{E}[N_k(T)]$ for any k with $\Delta_k > 0$.

Let $k^* \in \underset{k}{\arg\max}\ \mu_k$.

$$\mathbb{E}[N_k(T)] \leqslant \sum_{t=1}^{T} \mathbb{P}\left(\frac{1}{t} \sum_{s=1}^{t} X_k(r) - X_{k^*}(r) \geqslant 0 \right)$$

$$\leqslant \sum_{t=1}^{T} \mathbb{P}\left(\sum_{s=1}^{t} (X_k(r) - \mu_k) - \sum_{s=1}^{t} (X_{k^*}(r) - \mu_{k^*}) \geqslant t\Delta_k \right)$$

$$\leqslant \sum_{t=1}^{T} e^{-t\Delta_k^2} \leqslant \frac{e^{-\Delta_k^2}}{1 - e^{-\Delta_k^2}} = \frac{1}{e^{\Delta_k^2} - 1}$$

Hoeffding inequality.

$e^{\Delta_k^2} - 1 \geqslant \Delta_k^2$

$$\leqslant \frac{1}{\Delta_k^2}.$$

So $\mathbb{E}[R_T] = \sum_{k=1}^{K} \Delta_k \, \mathbb{E}[N_k(T)]$

$$\leqslant \sum_{k, \mu_k < \mu^*} \Delta_k \cdot \frac{1}{\Delta_k^2} \qquad \square$$

Bandit 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 $X_{a_t}(t) \in [0,1]$

$\mid a_t$ is $\sigma(U_0, X_{a_1}(1), U_1, \dots, X_{a_{t-1}}(t-1), U_{t-1})$-measurable

$$R_T = \max_{k \in [K]} \sum_{t=1}^{T} \mu_k - \sum_{t=1}^{T} \mu_{a_t}$$

$\longrightarrow$ only observe the reward of the pulled arm

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

This setting is sometimes called **random table model** and is known to be equivalent (from a probabilistic point of view) to the following **stack of rewards model**.

Bandit Setting (stack of rewards 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 $X_{a_t}(N_{a_t}(t-1)) \in [0,1]$

$\mid a_t$ is $\sigma(U_0, X_{a_1}(1), U_1, \dots, X_{a_{t-1}}(N_{a_{t-1}}(t-1)), U_{t-1})$-measurable

Same definition of regret

Stack of rewards model allows easier proofs, but heavier in notations.

Unless specified otherwise, we will consider the random table model in the following.

Notation

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

Random stack $\rightarrow \hat{\mu}_k(t) = \frac{1}{N_k(t)} \sum_{s=1}^{N_k(t)} X_k(s)$

Greedy algorithm (Bandit setting)

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

$a_t = t$

For $t \geq K+1$:

$a_t \in \operatorname{argmax}_{k \in [K]} \hat{\mu}_k(t-1)$

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

$$\mathbb{E}[R_T] \geq \frac{T-1}{32}.$$

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.

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, ie

$$a_t = \arg\max_k \hat{\mu}_k(nK)$$

Simple algorithm clearly separating exploration from exploitation.

Easy analysis

Theorem:

For any $1 \leq n \leq T/K$, 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 $n \leq T/K$, $\quad N_k(T) = \begin{cases} n & \text{if } k \neq \arg\max_{k'} \hat{\mu}_{k'}(nK) \\ n + (T-nK) & \text{if } k = \arg\max_{k'} \hat{\mu}_{k'}(nK) \end{cases}$

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

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

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

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

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

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

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 $n = \left\lceil \frac{\ln(T)}{\Delta^2} \right\rceil$

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

- Actually, we even showed a high probability regret bound, i.e. with $n = \left\lceil \frac{\ln(K/\delta)}{\Delta^2} \right\rceil$, ETC satisfies with probability at least $1-\delta$,

$$R_T \leq \left\lceil \frac{\ln(K/\delta)}{\Delta^2} \right\rceil \sum_{k=1}^{K} \Delta_k.$$

- ETC is easy to analyse $\rightarrow$ direct application of Hoeffding inequality

Yet, this use of Hoeffding ineq. is not always possible.

Instead, we use the following concentration lemma.

Lemma: (bandit concentration)

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

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

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

1) 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!

Hoeffding inequality indeed gives

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

But here, n is a random variable and is not independent from $\hat{\mu}_k(t)$

- What if instead we used Azuma-Hoeffding on $(X_k(s) - \mu_k)\mathbb{1}_{\{a_s = k\}}$?

↓

martingale increment bounded between $-\mu_k$ and $1-\mu_k$.

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

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

difference with our Lemma

getting rid of this $\sqrt{\frac{t}{N_k(t)}}$ factor is a big deal!

Proof 1 (stack of rewards model)

$$Z_n = \sum_{s=1}^{n} \left(X_k(s) - \mu_k\right)$$
$$\hat{\mu}_k(t) - \mu_k = \frac{1}{N_k(t)} Z_{N_k(t)}$$

Hoeffding inequality yields for any n.

$$\mathbb{P}\left(\frac{1}{n} Z_n \geq \sqrt{\frac{\ln(1/\delta)}{2n}}\right) \leq \delta$$

Moreover, $N_k(t) \in [\![0,t]\!]$ a.s., so we can do a union bound:

(bound is automatic for $n=0$)

$$\mathbb{P}\left(\exists n \in [\![1,t]\!],\ \frac{Z_n}{n} \geq \sqrt{\frac{\ln(1/\delta)}{n}}\right) \leq \delta t.$$

so that in particular $\mathbb{P}\left(\frac{Z_{N_k(t)}}{N_k(t)} \geq \sqrt{\frac{\ln(1/\delta)}{n}}\right) \leq \delta t.$ by inclusion of the event.

Symmetric arguments for the second inequality. □

Proof 2 (random table model)

Let $Z_t = \sum_{s=1}^{t} \left(X_k(s) - \mu_k\right) \mathbb{1}_{a_s = k}$.
$\hat{\mu}_k(t) - \mu_k = \frac{1}{N_k(t)} Z_t$

Here using a union bound on all the possible values of $(\mathbb{1}_{a_s=k})_{s=1\dots t}$ is a bad idea...

1) We first prove that $\forall x \in \mathbb{R},\ \mathbb{E}\left[e^{x Z_t - \frac{x^2}{8} N_k(t)}\right] \leq 1.$

For that, we show that $M_t = \exp\left(x Z_t - \frac{x^2}{8} N_k(t)\right)$ is a supermartingale, so that $\mathbb{E}[M_t] \leq \mathbb{E}[M_0] = 1.$

Let $\mathcal{F}_{t-1} = \sigma\left(U, X_{a_1}(1), \dots, X_{a_{t-1}}(t-1)\right)$

a_t is $\mathcal{F}_{t-1}$ measurable so that:

$$\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] = \mathbb{E}\left[e^{\left(x(X_k(t)-\mu_k)-\frac{x^2}{8}\right)\mathbb{1}_{a_t=k}} \middle| \mathcal{F}_{t-1}\right] M_{t-1}$$

$$= \left(\mathbb{E}\left[e^{x(X_k(t)-\mu_k)-\frac{x^2}{8}} \middle| \mathcal{F}_{t-1}\right]\mathbb{1}_{a_t=k} + \mathbb{1}_{a_t\neq k}\right) M_{t-1}$$

Hoeffding's lemma (conditional) gives $\ln\left(\mathbb{E}\left[e^{x(X_k(t)-\mu_k)} \mid \mathcal{F}_{t-1}\right]\right) \leqslant \frac{x^2}{8}$

so $\mathbb{E}\left[e^{x(X_k(t)-\mu_k)-\frac{x^2}{8}} \mid \mathcal{F}_{t-1}\right] \leqslant 1$.

i.e. $\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] \leqslant (\mathbb{1}_{a_t=k} + \mathbb{1}_{a_t\neq k}) M_{t-1}$

$\leqslant M_{t-1}$.

so we showed $\mathbb{E}\left[e^{xZ_t - \frac{x^2}{8}N_k(t)}\right] \leqslant 1$.

2) We now prove that $\forall \varepsilon > 0,\ \forall n \geqslant 1,$

$$\mathbb{P}(Z_t \geqslant \varepsilon \text{ and } N_k(t) = n) \leqslant e^{-\frac{2\varepsilon^2}{n}}$$

Indeed, by Markov-Chernoff bounding for any $x > 0$:

$$\mathbb{P}(Z_t \geqslant \varepsilon \text{ and } N_k(t) = n) \leqslant e^{-x\varepsilon}\, \mathbb{E}\left[e^{xZ_t}\mathbb{1}_{\{N_k(t)=n\}}\right]$$

$$= e^{-x\varepsilon + \frac{x^2}{8}n}\, \mathbb{E}\left[e^{xZ_t - \frac{x^2}{8}N_k(t)}\mathbb{1}_{\{N_k(t)=n\}}\right]$$

$$\leqslant e^{-x\varepsilon + \frac{x^2}{8}n}\, \underbrace{\mathbb{E}\left[e^{xZ_t - \frac{x^2}{8}N_k(t)}\right]}_{\leqslant 1 \text{ thanks to } 1)}$$

$$\leqslant e^{-x\varepsilon + \frac{x^2}{8}n}$$

Taking $x = \frac{4\varepsilon}{n}$ finally yields

$$\boxed{\mathbb{P}(Z_t \geqslant \varepsilon \text{ and } N_k(t) = n) \leqslant e^{-\frac{2\varepsilon^2}{n}}}.$$

$(\#)$

3) We conclude using a union bound: (as in the stack of rewards model)

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

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

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

$$\leq \sum_{n=1}^{t} e^{-\ln(1/\delta)} = t\,\delta.$$

$\boxtimes$

Notes on this proof:

- We saw last week that the conditional version of Hoeffding's lemma could be generalized into

X bounded random variable, U, V two $\mathcal{G}$-measurable random variables with $U \leq X \leq V$ a.s.

then $\forall \eta \in \mathbb{R}$,

$$\ln \mathbb{E}[e^{\eta X}|\mathcal{G}] \leq \eta\, \mathbb{E}[X|\mathcal{G}] + \frac{\eta^2}{8}(V-U)^2$$

This can be applied to

$$Z_t = (X_k(t)-\mu_k)\, \mathbb{1}_{\{a_t=k\}}$$

$$\mathcal{G} = \mathcal{F}_{t-1}$$

$$U_t = -\mu_k\, \mathbb{1}_{\{a_t=k\}}$$

$$V_t = (1-\mu_k)\, \mathbb{1}_{\{a_t=k\}}$$

and directly entails $\mathbb{E}\left[e^{\eta(X_k(t)-\mu_k)\mathbb{1}_{\{a_t=k\}}} \,\middle|\, \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\eta^2}{8}\mathbb{1}_{\{a_t=k\}}\right)$ without the need for the

$1 = \mathbb{1}_{\{a_t=k\}} + \mathbb{1}_{\{a_t\neq k\}}$ trick used in step 1).

- The question is: Don't we have a generalized version of the Hoeffding-Azuma inequality with such predictable ranges $V_t - U_t$?

Yes, we do have something in terms of constant upper bounds $V_t - U_t \leqslant \Delta_t \in \mathbb{R}$ a.s.

but $V_t - U_t = \mathbb{1}_{\{a_t = k\}}$ can only be bounded by $\Delta_t = 1$ here, so steps 2) and 3) are still needed.

- For unbounded, but σ-sub-Gaussian variables $X_k(t)$, we still have, thanks to the stack of rewards model: $\mathbb{P}\left(\mu_k - \hat{\mu}_k(t) \geqslant \sqrt{\frac{2\ln(1/\delta)}{\sigma^2 N_k(t)}}\right) \leqslant t\delta$.

ε-Greedy

sequence of probabilities ε_t.

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

$a_t = t$

For $t \geqslant K+1$:

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

explore uniformly at random

Theorem

For $\varepsilon_t = \min\left\{1, \frac{cK}{t\Delta^2}\right\}$ where c is a large enough universal constant, ε-greedy satisfies for a large enough universal constant c'

$$R_T \leqslant \frac{c'}{\Delta^2} \sum_{k=1}^{K} \left(\Delta_k \ln T + 1\right)$$

Proof: For any k with $\Delta_k > 0$,

$$\mathbb{P}(a_t = k) \leq \frac{\varepsilon_t}{K} + \mathbb{P}\left(\hat{\mu}_k(t-1) \geq \hat{\mu}_{k^*}(t-1)\right)$$

$$\leq \frac{\varepsilon_t}{K} + \mathbb{P}\left(\hat{\mu}_k(t-1) - \mu_k \geq \frac{\Delta_k}{2}\right) + \mathbb{P}\left(\mu_{k^*} - \hat{\mu}_{k^*}(t-1) \geq \frac{\Delta_k}{2}\right).$$

$$\mathbb{P}\left(\hat{\mu}_k(t-1) - \mu_k \geq \frac{\Delta_k}{2}\right) \leq \sum_{n=1}^{\lfloor x_t \rfloor} \mathbb{P}(N_k(t-1) \leq x_t) + \sum_{n=\lfloor x_t \rfloor+1}^{t-1} \mathbb{P}\left(\hat{\mu}_k(t-1) - \mu_k \geq \frac{\Delta_k}{2} \text{ and } N_k(t-1) = n\right)$$ where $0 \leq x_t \leq t-1$

$$\leq \mathbb{P}(N_k(t-1) \leq x_t) + \sum_{n=\lfloor x_t \rfloor+1}^{t-1} e^{-\frac{\Delta_k^2}{2} n} \quad (*)$$

$$\leq \mathbb{P}\big(\underbrace{N_k^R}_{\text{number of times } k \text{ is pulled at random (i.e. following the } \varepsilon \text{ prob. event)}}(t-1) \leq x_t\big) + 2\frac{e^{-\frac{\Delta_k^2}{2} x_t}}{\Delta_k^2}$$

$$\mathbb{E}[N_k^R(t-1)] = 1 + \frac{1}{K}\sum_{s=K+1}^{t-1} \varepsilon_s = \frac{1}{K}\sum_{s=1}^{t-1} \varepsilon_s.$$

$$\mathrm{Var}(N_k^R(t-1)) = \sum_{s=K+1}^{t-1} \frac{\varepsilon_s}{K}\left(1 - \frac{\varepsilon_s}{K}\right) \leq \frac{1}{K}\sum_{s=1}^{t-1} \varepsilon_s$$

Recall

Bernstein Inequality

lecture #2

Let $X_1, \dots, X_t$ be random variables in $[0,1]$ s.t. $\mathrm{Var}[X_s | X_1, \dots, X_{s-1}] = \sigma_t^2$.

Then for all $\varepsilon > 0$:
$$\mathbb{P}\left(\sum_{s=1}^{t} X_s - \mathbb{E}[X_s | X_1, \dots, X_{s-1}] \leq -\varepsilon\right) \leq \exp\left(\frac{-\varepsilon^2/2}{\sum_{s=1}^{t} \sigma_s^2 + \frac{\varepsilon}{2}}\right)$$

So here for $x_t = \frac{1}{2K}\sum_{s=1}^{t-1} \varepsilon_s$

$$\mathbb{P}(N_k^R(t-1) \leq x_t) = \mathbb{P}\left(N_k^R(t-1) - \mathbb{E}[N_k^R(t-1)] \leq -x_t \right)$$

$$\leq \exp\left(- \frac{x_t^2/2}{\frac{5}{2} x_t} \right) = e^{-\frac{x_t}{5}}.$$

Moreover: $x_t = \frac{1}{2K} \sum_{s=1}^{t-1} \min\left(1, \frac{cK}{s\Delta^2} \right)$

For $t \geq \left\lfloor \frac{\Delta^2}{cK} \right\rfloor + 1$, $x_t = \left\lfloor \frac{cK}{\Delta^2} \right\rfloor \cdot \frac{1}{2K} + \frac{1}{2K} \sum_{s=\lfloor \frac{cK}{\Delta^2} \rfloor + 1}^{t-1} \frac{cK}{s\Delta^2}$

$$\geq \left(\frac{c}{2\Delta^2} - \frac{\Delta^2/K}{2\Delta^2} \right) + \frac{c}{2\Delta^2} \ln\left(\frac{t-1}{\lfloor \frac{cK}{\Delta^2} \rfloor} \right)$$

$\sum_{s=a}^{b} \frac{1}{s} \geq \int_{a-1}^{b} \frac{1}{s} ds$

$$x_t \geq \frac{c-1}{2\Delta^2} \ln\left(\frac{e(t-1)\Delta^2}{cK} \right)$$

Recap: $\mathbb{P}(a_t = k) \leq \frac{\varepsilon_t}{K} + \mathbb{P}\left(\hat{\mu}_k(t-1) - \mu_k \geq \frac{\Delta_k}{2} \right) + \mathbb{P}\left(\mu_{k^*} - \hat{\mu}_{k^*}(t-1) \geq \frac{\Delta_k}{2} \right).$

with

$$\mathbb{P}\left(\hat{\mu}_k(t-1) - \mu_k \geq \frac{\Delta_k}{2} \right) \leq e^{-\frac{x_t}{5}} + \frac{2e^{-\frac{\Delta_k^2}{2} x_t}}{\Delta_k^2}$$

So $\mathbb{P}(a_t = k) \leq \frac{c}{\Delta^2 t} + 2e^{-x_t/5} + \frac{4}{\Delta_k^2} e^{-\Delta_k^2 x_t/2}$

with $x_t \geq \frac{c-1}{2\Delta^2} \ln\left(\frac{e(t-1)\Delta^2}{cK} \right)$

This yields for a large enough c and $t \geq \lfloor \frac{\Delta^2}{cK} \rfloor + 1$,

$$\mathbb{P}(a_t = k) = O\left(\frac{1}{\Delta^2 t}\right)$$ so that for large enough c', c:

so that:

$$\sum_{t=\lfloor \frac{\Delta^2}{cK} \rfloor + 1}^{T} \mathbb{P}(a_t = k) \leq \frac{c'}{\Delta^2} \ln(T)$$

hence regret is bounded as:

$$R_T \leq \frac{cK}{\Delta^2} + \sum_{k=1}^{K} \frac{c' \Delta_k}{\Delta^2} \ln(T)$$

$$\leq \frac{c'}{\Delta^2} \sum_{k=1}^{K} (\Delta_k \ln(T) + 1) \qquad \square$$

does not depend on any parameters $(\mu, K, \Delta, T \ldots)$

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 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 these methods: (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.

Moreover, ETC needs knowledge of T.

$\rightarrow$ if T is unknown, we can use the doubling trick (for any algo)