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:

$$\bar{R}_T = \max_{k \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} X_k(t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} X_{a_t}(t)\right]$$

↳ expectation wrt the realizations of $X(t)$ and eventually the agent policy

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?

$$\boxed{\bar{R}_T = T \max_k \mu_k - \mathbb{E}\left[\sum_{t=1}^{T} X_{a_t}(t)\right]}$$

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

Notations:

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

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

Lemma:

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

Proof:

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \mu^* - X_{a_t}(t)\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} \mu^* - \sum_{k=1}^{K} \mathbb{1}_{a_t=k} X_k(t)\right]$$

$$= \sum_{t=1}^{T} \sum_{k=1}^{K} \mathbb{E}\left[(\mu^* - X_k(t)) \mathbb{1}_{a_t=k}\right]$$

$$= \sum_{k=1}^{K} \sum_{t=1}^{T} (\mu^* - \mu_k) \mathbb{E}\left[\mathbb{1}_{a_t=k}\right] \qquad X_k(t) \perp\!\!\!\perp a_t$$

$$= \sum_{k=1}^{K} \Delta_k \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}_{a_t=k}\right]$$

$$= \sum_{k=1}^{K} \Delta_k \mathbb{E}[N_k(T)] \qquad \square$$

Greedy algorithm (or Follow The Leader)

Choose a_1 arbitrarily

For $t \geq 2$:

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

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

$$R_T \leq \sum_{k, \Delta_k > 0} \frac{1}{\Delta_k}$$

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

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

Let $k^* \in \underset{k}{\operatorname{argmax}} \, \mu_k$.

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

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

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

Hoeffding inequality.

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

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

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

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

Bandit Setting

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

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

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

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

→ 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

Notation

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

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:

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

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.

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 $1-\mu_k$.

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

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

difference with our Lemma

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

Proof: Let $Z_t = \sum_{s=1}^{t}(X_k(s)-\mu_k)\mathbb{1}_{a_s=k}$.

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

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

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:

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

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

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

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

3) We conclude using a union bound:

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

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

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

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

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 $V_t - U_t \leq \Delta_t \in \mathbb{R}$ 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: $\mathbb{P}\left(\mu_k - \hat{\mu}_k(t) \geq \sqrt{\frac{2\ln(1/\delta)}{\sigma^2 N_k(t)}}\right) \leq t\delta$.

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

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'

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

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

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

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

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

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

where $0 \leq x_t \leq t-1$

$$\leq \mathbb{P}(N_k(t-1) \leq x_t) + \sum_{n=\lfloor x_t \rfloor + 1}^{t-1} e^{-\frac{x \Delta_k}{2} n + \frac{x^2}{8} n} \, \mathbb{E}\left[e^{x Z_{t-1} - \frac{x^2}{8} N_k(t-1)} \mathbb{1}_{\{N_k(t-1) = n\}}\right]$$

for any $x > 0$

similarly to previous proof with $Z_{t-1} = \sum_{s=1}^{t-1} (X_k(s) - \mu_k) \mathbb{1}_{\{a_s = k\}}$

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

$\mathbb{E}\left[e^{x Z_{t-1} - \frac{x^2}{8} N_k(t-1)}\right] \leq 1$ and taking $x = 2\Delta_k$

$$\leq \mathbb{P}\Big(\underbrace{N_k^R(t-1)}_{\text{number of times } k \text{ is pulled at random}} \leq x_t\Big) + 2 \frac{e^{-\frac{\Delta_k^2}{2} x_t}}{\Delta_k^2}$$

(i.e. following the ε prob. event)

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

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

lecture #2

Recall

Bernstein Inequality

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

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

so here for $x_r = \frac{1}{2K} \sum_{s=1}^{r-1} \varepsilon_s$

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

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

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

For $r \geq \left\lfloor \frac{\Delta^2}{cK} \right\rfloor + 1$: $x_r = \left\lfloor \frac{cK}{\Delta^2} \right\rfloor \cdot \frac{1}{2K} + \frac{1}{2K} \sum_{s=\lfloor \frac{cK}{\Delta^2} \rfloor + 1}^{r-1} \frac{cK}{s\Delta^2}$

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

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

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

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

with

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

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

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

$$\leq \frac{c}{\Delta^2 t} + 2\left(\frac{cK}{e(t-1)\Delta^2}\right)^{\frac{c-1}{10\Delta^2}}$$

$\geq \frac{c-1}{10}$

$$+ \frac{4}{\Delta_k^2}\left(\frac{cK}{e(t-1)\Delta^2}\right)^{\frac{c-1}{4}\frac{\Delta_k^2}{\Delta^2}}$$

for $c \geq 11$: $$\mathbb{P}(a_t = k) \leq \frac{c}{\Delta^2 t} + \frac{c'}{\Delta^2} \frac{\ln(a(t-1))}{(a(t-1))^2}$$

$a = \frac{e\Delta^2}{cK}$

for some constant c'.

$$\sum_{t=\lceil\frac{1}{a}\rceil}^{T} \mathbb{P}(a_t = k) \leq \frac{c_1}{\Delta^2}\ln(T) + \frac{c_2}{\Delta^2}$$

So that for a universal constant c'' large enough

↳ does not depend on any parameter K, T, μ, Δ, etc.

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

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

$\blacksquare$

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:

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

see exercise session #2

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