More on "adversarial bandits"

Adversarial bandits: high-probability regret bounds (A brief and suboptimal view...)

Setting (reminder): at each round $t=1,2,\dots$

1. the opponent and the decision-maker simultaneously choose $\ell_t = (\ell_{jt})_{j=1,\dots,N}$ with $\ell_{jt} \in [0,M]$ and $I_t \sim p_t$, where $p_t \in \mathcal{P}(\{1,\dots,N\})$;
2. the opponent gets to see p_t and I_t; the decision-maker only observes $\ell_{I_t t}$ (her own loss).

↳ She wants to control her regret

$$R_T = \sum_{t=1}^{T} \ell_{I_t t} - \min_{j=1,\dots,N} \sum_{t=1}^{T} \ell_{jt}$$

She resorts to the (conditionally) unbiased estimators

$$\hat{\ell}_{jt} = \frac{\ell_{I_t t}}{p_{jt}} \mathbb{1}_{\{I_t = j\}}$$

Denoting by $\mathcal{F}_{t-1} = \sigma(p_s, I_s, \ell_s, s \le t-1)$ and possibly some auxiliary randomizations and $\mathcal{F}_0 = \{\emptyset, \Omega\}$ we showed:

$$\forall t \ge 1,\ \forall j \in \{1,\dots,N\},\quad \mathbb{E}[\hat{\ell}_{jt} \mid \mathcal{F}_{t-1}] = \ell_{jt}$$

Main technical ingredient needed in the proof:

Be able to relate $\sum_{t=1}^{T} \hat{\ell}_{jt}$ to $\sum_{t=1}^{T} \ell_{jt}$ with high probability. We will use martingale inequalities but will need a control from below on the p_{jt}. Hence the Exp3 algorithm:

$$\text{for } t \ge 2,\ \forall j,\quad p_{jt} = (1-\gamma_t) \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \hat{\ell}_{js}\right)}{\sum_{k=1}^{N} \exp\left(-\eta_t \sum_{s=1}^{t-1} \hat{\ell}_{ks}\right)} + \frac{\gamma_t}{N}$$

↑ exploitation term ↑ exploration term

Hence the name Exp3 for: exponential weights for exploration and exploitation

Then: $\hat{\ell}_{jt} \in \left[0, \frac{MN}{\gamma_t}\right]$ as $p_{jt} \geq \gamma_t/N$ and $\ell_{I_t,t} \in [0,M]$

The Hoeffding-Azuma inequality ensures that with probability at least $1-\delta$,

$$(*) \qquad \sum_{t=1}^{T} \hat{\ell}_{jt} \leq \sum_{t=1}^{T} \underbrace{\mathbb{E}\left[\hat{\ell}_{jt} \mid \mathcal{F}_{t-1}\right]}_{=\ell_{jt}} + MN \underbrace{\sqrt{\sum_{t=1}^{T} \frac{1}{2\gamma_t^2} \ln\frac{1}{\delta}}}_{\text{depends on the range}}$$

A sharper bound is in terms of the conditional variances:

$$\mathrm{Var}_{\mathcal{F}_{t-1}}\left(\hat{\ell}_{jt}\right) \leq \mathbb{E}\left[\hat{\ell}_{jt}^2 \mid \mathcal{F}_{t-1}\right] \leq M^2\, \mathbb{E}\left[\mathbb{1}_{I_t=j}/p_{jt}^2 \mid \mathcal{F}_{t-1}\right] = M^2/p_{jt} \leq \frac{M^2 N}{\gamma_t}$$

(as $\ell_{I_t,t}^2 \leq M^2$)

↳ Bernstein's inequality for martingales ensures that with probability at least $1-\delta$,

$$(**) \qquad \sum_{t=1}^{T} \hat{\ell}_{jt} \leq \sum_{t=1}^{T} \ell_{jt} + O\left(M\sqrt{N}\sqrt{\sum_{t=1}^{T} \frac{1}{\gamma_t} \ln\frac{1}{\delta}}\right)$$

we will gain orders of magnitude as γ_t will be of the form $t^{-\alpha}$

→ With (*) we would get a regret bound of order $T^{3/4}$

while with (**) we can get the desired $\sqrt{T}$, after an additional twist (we will first exhibit a $T^{2/3}$ bound and only hint at the $\sqrt{T}$ rate as an exercise).

Bernstein's inequality for martingales.

We proved Bernstein's lemma with true expectations but given its proof (where we used only the monotonicity of $\mathbb{E}$), it appears that we can replace all $\mathbb{E}$ by $\mathbb{E}[\,\cdot\,|\mathcal{G}_j]$.

But I realized meanwhile that my proof was suboptimal as it used a lower bound m on the random variable X at hand, which is an unconvenient assumption. Let's re-do it.

Lemma: Let X be a random variable with $X - \mathbb{E}[X|\mathcal{G}_j] \leq M$ and $\mathcal{G}_j$ a σ-algebra; then:

$$\forall \eta \geq 0, \quad \ln \mathbb{E}[e^{\eta X}|\mathcal{G}_j] \leq \eta\, \mathbb{E}[X|\mathcal{G}_j] + \frac{e^{\eta M} - \eta M - 1}{M^2} \mathrm{Var}_{\mathcal{G}_j}(X)$$

or put differently,

$$\mathbb{E}\left[e^{\eta(X - \mathbb{E}[X|\mathcal{G}_j])} \,|\, \mathcal{G}_j\right] \leq \exp\left(\frac{e^{\eta M} - \eta M - 1}{M^2} \mathrm{Var}_{\mathcal{G}_j}(X)\right)$$

($\mathrm{Var}_{\mathcal{G}_j}(X)$: conditional variance of X)

Proof: (sketch) $\eta(X - \mathbb{E}[X|\mathcal{G}_j]) \leq \eta M$ and $x \in \mathbb{R} \longmapsto \frac{e^x - x - 1}{x^2}$ is increasing,

so that $e^{\eta(X - \mathbb{E}[X|\mathcal{G}_j])} - \eta(X - \mathbb{E}[X|\mathcal{G}_j]) - 1 \leq \eta^2 (X - \mathbb{E}[X|\mathcal{G}_j])^2 \frac{e^{\eta M} - \eta M - 1}{\eta^2 M^2}$

Taking $\mathbb{E}[\,\cdot\,|\mathcal{G}_j]$:

$$\mathbb{E}\left[e^{\eta(X - \mathbb{E}[X|\mathcal{G}_j])} \,|\, \mathcal{G}_j\right] - 1 \leq \mathrm{Var}_{\mathcal{G}_j}(X) \frac{e^{\eta M} - \eta M - 1}{M^2}$$

$\ln u \overset{\text{as}}{\leq} u - 1$ $\forall u > 0$ $\hookrightarrow$ $\geq \ln \mathbb{E}\left[e^{\eta(X - \mathbb{E}[X|\mathcal{G}_j])} \,|\, \mathcal{G}_j\right]$, which concludes the proof.

Theorem: Let $(\mathcal{F}_t)_{t \geq 0}$ be a filtration and let $(X_t)_{t \geq 1}$ be a sequence of adapted random variables, with $X_t - \mathbb{E}[X_t | \mathcal{F}_{t-1}] \leq M$ a.s., $\forall t$.

- probabilistic version: $\forall \varepsilon > 0$, $\forall V > 0$,

$$\mathbb{P}\left\{ \sum_{t=1}^{T} X_t - \sum_{t=1}^{T} \mathbb{E}[X_t | \mathcal{F}_{t-1}] \geq \varepsilon \text{ and } \sum_{t=1}^{T} \mathrm{Var}_{\mathcal{F}_{t-1}}(X_t) \leq V \right\} \leq \exp\left(-\frac{\varepsilon^2}{2V + \frac{2}{3} M \varepsilon}\right)$$

- Statistical version:

$$\mathbb{P}\left\{ \sum_{t=1}^{T} X_t - \sum_{t=1}^{T} \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \geqslant \sqrt{2V \ln\frac{1}{\delta}} + \frac{2}{3} M \ln\frac{1}{\delta} \right\} \leqslant \delta$$

where V is a numerical constant >0 and $\sum_{t=1}^{T} \mathrm{Var}_{\mathcal{F}_{t-1}}(X_t) \leqslant V$

Proof: $\eta > 0$,

$$M_t = \exp\left(\eta \sum_{s=1}^{t} \left(X_s - \mathbb{E}[X_s \mid \mathcal{F}_{s-1}] \right) - \left(e^{\eta M} - \eta M - 1 \right)/M^2 \sum_{s=1}^{t} \mathrm{Var}_{\mathcal{F}_{s-1}}(X_s) \right)$$

by Bernstein's lemma, $(M_t)_t$ is an $(\mathcal{F}_t)_t$-adapted supermartingale.

Thus

$$\mathbb{P}\left\{ \sum_{t=1}^{T} \underbrace{\left(X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \right)}_{\overset{\text{not.}}{=} S_T} \geqslant \varepsilon \quad \text{and} \quad \underbrace{\sum_{t=1}^{T} \mathrm{Var}_{\mathcal{F}_{t-1}}(X_t)}_{\overset{\text{not.}}{=} \sigma_T^2} \leqslant V \right\}$$

$\forall \eta > 0$

$$= \mathbb{P}\left\{ e^{\eta S_T - \left(e^{\eta M} - \eta M - 1\right)/M^2 \sigma_T^2} \geqslant e^{\eta\varepsilon - \left(e^{\eta M} - \eta M - 1\right)/M^2 \sigma_T^2} \text{ and } \sigma_T^2 \leqslant V \right\}$$

($e^{\eta S_T - \left(e^{\eta M} - \eta M - 1\right)/M^2 \sigma_T^2} = M_T$; $\sigma_T^2 \leqslant V$)

$$\leqslant \mathbb{P}\left\{ M_T \geqslant \exp\left(\eta\varepsilon - \left(e^{\eta M} - \eta M - 1\right)/M^2 V \right) \right\}$$

$$\leqslant \underbrace{\exp\left(-\eta\varepsilon + \left(e^{\eta M} - \eta M - 1\right)/M^2 V \right)}_{\text{to be optimized over } \eta > 0:} \underbrace{\mathbb{E}[M_T]}_{\leqslant \mathbb{E}M_0 = 1 \text{ as } (M_t)_t \text{ is a supermartingale}}$$

(Markov's ineq.)

$$f(x) = -x\varepsilon + \frac{V}{M}\left(e^x - x - 1\right) \qquad \text{where } \eta M = x$$

$$f'(x) = -\varepsilon + \frac{V}{M}\left(e^x - 1\right)$$

$$f''(x) = \frac{V e^x}{M} \geqslant 0$$

unique minimizer x^* s.t. $f'(x^*) = 0$, i.e. $e^{x^*} = 1 + \frac{M\varepsilon}{V}$

$$x^* = \ln\left(1 + \frac{M\varepsilon}{V}\right)$$

$$\hookrightarrow \eta^* = \frac{1}{M}\ln\left(1 + \frac{M\varepsilon}{V}\right)$$

The bound is

$$\exp\left(-\frac{\varepsilon}{M}\ln\left(1 + \frac{M\varepsilon}{V}\right) + \frac{V}{M^2}\left(\left(1 + \frac{M\varepsilon}{V}\right) - \ln\left(1 + \frac{M\varepsilon}{V}\right) - 1 \right) \right)$$

$$= \exp\left(\frac{\varepsilon}{M} - \frac{\varepsilon}{M}\ln\left(1 + \frac{M\varepsilon}{V}\right) - \frac{V}{M^2}\ln\left(1 + \frac{M\varepsilon}{V}\right) \right)$$

$$= \exp\left(-\frac{V}{M^2}\left(-\frac{M\varepsilon}{V} + \left(\frac{M\varepsilon}{V}+1\right)\ln\left(1+\frac{M\varepsilon}{V}\right)\right)\right)$$

$$= \exp\left(-\frac{V}{M^2} h\left(\frac{M\varepsilon}{V}\right)\right) \qquad \text{where} \quad h(u) = (1+u)\ln(1+u) - u$$

We conclude by noting that $\forall u \in \mathbb{R}$, $h(u) \geq \frac{u^2}{2+\frac{2}{3}u}$:

$$\frac{V}{M^2} h\left(\frac{M\varepsilon}{V}\right) \geq \frac{V}{M^2}\left(\frac{M\varepsilon}{V}\right)^2 \Big/ 2+\frac{2}{3}\frac{M\varepsilon}{V}$$

$$= \varepsilon^2 \Big/ V\left(2+\frac{2}{3}\frac{M\varepsilon}{V}\right).$$

probabilistic version

Statistical version

↓

It suffices to show that for $\varepsilon = \sqrt{2V\ln\frac{1}{\delta}} + \frac{2}{3}M\ln\frac{1}{\delta}$, we have $\exp\left(-\varepsilon^2/(2V+\frac{2}{3}M\varepsilon)\right) \leq \delta$.

Indeed, $\varepsilon^2 = \left(\sqrt{2V\ln\frac{1}{\delta}} + \frac{2}{3}M\ln\frac{1}{\delta}\right)\varepsilon$ where $\varepsilon \geq \sqrt{2V\ln 1/\delta}$

$$\geq 2V\ln\frac{1}{\delta} + \frac{2}{3}M\varepsilon\ln 1/\delta$$

$$= \left(2V + \frac{2}{3}M\varepsilon\right)\ln 1/\delta.$$

Application : Performance bound for Exp3

Theorem. For well-chosen sequences of $\eta_t \searrow$ and $\gamma_t \searrow$, we have, with probability $\geq 1-\delta$,

$$R_T = \sum_{t=1}^{T} \ell_{I_t t} - \min_{i=1..N} \sum_{t=1}^{T} \ell_{it} \leq C\left(T^{2/3} \ln \frac{N}{\delta}\right)$$

Remarks:
- Not quite the $\sqrt{T}$ rate we wanted! ↳ see Exercise to see how to correct this.
- It would be even worse with Hoeffding-Azuma ($T^{3/4}$ rate).
- Main issue : the deviation term $\sum \hat{\ell}_{it} - \sum \ell_{it} \leq \dots$ → to be improved.

Proof: Let $\tilde{p}_{jt} = \exp\left(-\eta_t \sum_{s=1}^{t-1} \hat{\ell}_{js}\right) \Big/ \sum_{k=1}^{N} \exp\left(-\eta_t \sum_{s=1}^{t-1} \hat{\ell}_{ks}\right)$

(so that $p_{jt} = (1-\gamma_t)\tilde{p}_{jt} + \gamma_t/N$).

A lemma used already in the proof of the expected bound shows that

$$\sum_{t=1}^{T} \sum_{j=1}^{N} \tilde{p}_{jt} \hat{\ell}_{jt} - \min_{i=1..N} \sum_{t=1}^{T} \hat{\ell}_{it} \leq \frac{\ln N}{\eta_T} + \sum_{t=1}^{T} \frac{\eta_t}{2} \sum_j \tilde{p}_{jt} \hat{\ell}_{jt}^2$$

thus $\sum_{t=1}^{T} \sum_{j=1}^{N} \left(\underbrace{(1-\gamma_t)}_{\leq 1} \tilde{p}_{jt} + \frac{\gamma_t}{N} \right) \hat{\ell}_{jt} - \min_{i=1..N} \sum_{t=1}^{T} \hat{\ell}_{it}$

$$\leq \sum_{t=1}^{T} \frac{\gamma_t}{N} \sum_j \hat{\ell}_{jt} + \left(\frac{\ln N}{\eta_T} + \sum_{t=1}^{T} \frac{\eta_t}{2} \sum_j \tilde{p}_{jt} \hat{\ell}_{jt}^2 \right)$$

↑ $(1-\gamma_t)\tilde{p}_{jt} \leq p_{jt}$

Therefore,

$$\sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \hat{\ell}_{jt} - \min_{i=1..N} \sum_{t=1}^{T} \hat{\ell}_{it} \leq \frac{\ln N}{\eta_T} + \frac{1}{2} \sum_{t=1}^{T} \frac{\eta_t}{1-\gamma_t} \sum_j p_{jt} \hat{\ell}_{jt}^2 + \sum_{t=1}^{T} \frac{\gamma_t}{N} \sum_j \hat{\ell}_{jt}$$

We already saw that: $\sum_j p_{jt} \hat{\ell}_{jt} = \ell_{I_t t}$ and that $\sum_j p_{jt} \hat{\ell}_{jt}^2 \leq M^2 \frac{\mathbb{1}_{I_t = j}}{p_{jt}}$

Bernstein's inequality applied $3N$ times:

$$\forall c,\ \sum_{t=1}^{T} \hat{\ell}_{it} \le \sum_{t=1}^{T} \ell_{it} + M\sqrt{N}\sqrt{\sum_{t=1}^{T} \frac{2}{\gamma_t} \ln\frac{3N}{\delta}} + \frac{2}{3}\frac{MN}{\gamma_T}\ln\frac{3N}{\delta}$$

$$\forall j,\ \sum_{t=1}^{T} \frac{\eta_t}{1-\gamma_t} \frac{\mathbb{1}_{I_t=j}}{p_{jt}} \le \sum_{t=1}^{T} \frac{\eta_t}{1-\gamma_t} + \sqrt{2\sum_{t=1}^{T} \frac{\eta_t^2}{(1-\gamma_t)^2}\frac{N}{\gamma_t}\ln\frac{3N}{\delta}} + \frac{2}{3}\max_{t\le T}\frac{\eta_t N}{(1-\gamma_t)\gamma_t}\ln\frac{3N}{\delta}$$

($\frac{\mathbb{1}_{I_t=j}}{p_{jt}}$: conditional variance $\le 1/p_{jt} \le N/\gamma_t$)

$$\forall j,\ \sum_{t=1}^{T} \gamma_t \hat{\ell}_{jt} \le \sum_{t=1}^{T} \gamma_t \ell_{jt} + \sqrt{2\sum_{t=1}^{T}\gamma_t M^2 N \ln\frac{3N}{\delta}} + \frac{2}{3}MN\ln\frac{3N}{\delta}$$

($\hat{\ell}_{jt}$: conditional variance $\le \gamma_t^2 M^2 N/\gamma_t = M^2 N\gamma_t$; $\ell_{jt} \le M$)

↑ All inequalities holding at the same time with probability at least $1-\delta$ (by the union bound).

The regret bound is of the form: with probability $\ge 1-\delta$,

$$R_T = \sum_{t=1}^{T} \ell_{I_t,t} - \min_{i=1,\dots,N}\sum_{t=1}^{T}\ell_{it}$$

$$\le \frac{\ln N}{\eta_T} + M\sqrt{N}\sqrt{\sum_{t=1}^{T}\frac{2}{\gamma_t}\ln\frac{3N}{\delta}} + M\sum_{t=1}^{T}\gamma_t + \frac{M^2}{2}\sum_{t=1}^{T}\frac{\eta_t}{1-\gamma_t}$$

[+ MANY OTHER TERMS]

but looking only at the γ_t we see the issue:

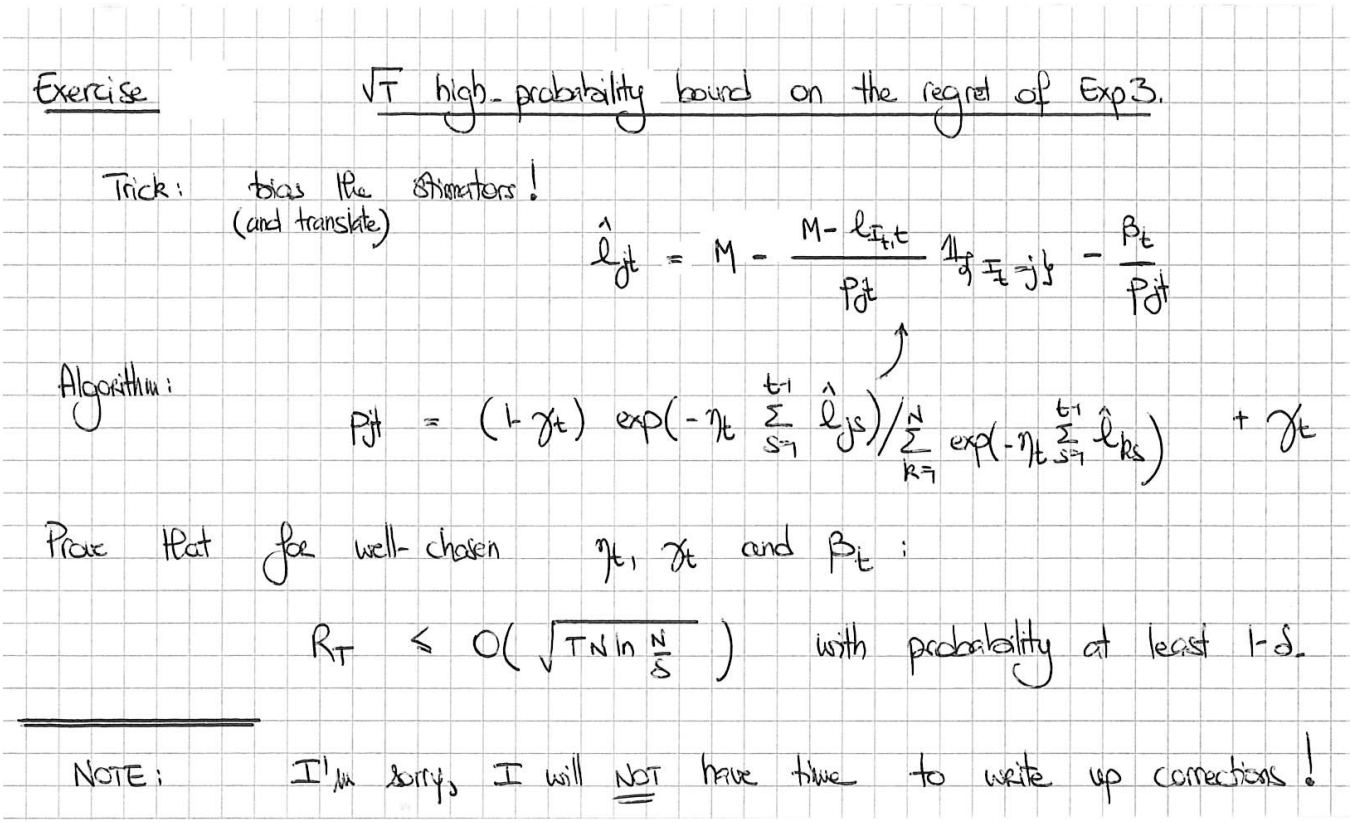
$\gamma_t \sim t^a$ ↪ $\left(\sum 1/\gamma_t\right)^{1/2} = O\left(t^{(-a+1)/2}\right)$

$\sum \gamma_t = O(t^{a+1})$

Choose a s.t. $-a+1 = 2(a+1)$ i.e., $a = -1/3$

and get a $T^{2/3}$ rate...

↳ You can show that indeed, the $T^{2/3}\ln\frac{N}{\delta}$ rate is achievable here.


Exercise
$\sqrt{T}$ high-probability bound on the regret of Exp3.
Trick: bias the estimators!
(and translate)
$\hat{\ell}_{jt} = M - \frac{M - \ell_{I_t,t}}{p_{jt}} \mathbb{1}_{I_t = j} - \frac{\beta_t}{p_{jt}}$
Algorithm:
$p_{jt} = (1-\gamma_t) \exp\left(-\eta_t \sum_{s=1}^{t-1} \hat{\ell}_{js}\right) \Big/ \sum_{k=1}^{N} \exp\left(-\eta_t \sum_{s=1}^{t-1} \hat{\ell}_{ks}\right) + \gamma_t$
Prove that for well-chosen η_t, γ_t and β_t:
$R_T \leq O\left(\sqrt{TN \ln \frac{N}{\delta}}\right)$ with probability at least $1-\delta$.
NOTE: I'm sorry, I will NOT have time to write up corrections!

Conclusion of this series of lectures

Links between:
- Distribution-free regret bounds for stochastic bandits
- Adversarial bandits

To that end we restrict our attention to the model $\mathcal{D} = \mathcal{P}([0,1])$, the set of all probability distributions over $[0,1]$.

Stochastic bandits

With each arm a is associated $\nu_a \in \mathcal{P}([0,1])$

For $t = 1, 2, \dots$
- The decision maker picks $I_t \in \{1..K\}$
- Her reward Y_t, which is such that $Y_t \mid I_t \sim \nu_{I_t}$, is her only piece of information

Aim: control the regret

$$\bar{R}_T = \max_{a=1..K} \mathbb{E}(\nu_a) - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]$$

Adversarial bandits

An opponent selects the payoffs g_{it}

For $t = 1, 2, \dots$
- The opponent picks (while, simultaneously) $(g_{1t} \dots g_{Kt}) \in [0,1]^K$
- The decision-maker picks $I_t \in \{1..K\}$
- Her payoff is $g_{I_t t}$ and this is the only piece of information she gets

Aim: control the regret

$$R_T = \max_{k=1..K} \sum_{t=1}^{T} g_{kt} - \sum_{t=1}^{T} g_{I_t t}$$

Typical adversarial results (Auer, Cesa-Bianchi, Freund, Schapire, 2002, later improved by Audibert and Bubeck, 2009):

Strategies such that for all opponents picking gains in $[0,1]$, for all $T \geq 1$,

with probability at least $1-\delta$, $R_T \leq C\sqrt{TK\ln(K/\delta)}$

$\mathbb{E}[R_T] \leq C\sqrt{TK\ln K}$

for some numerical constant C

where the probability and $\mathbb{E}$ are wrt decision-maker's internal randomization

For "oblivious" opponents (ie, when the g_{kt} do not "react" to the decision-maker's actions): the $\sqrt{\ln K}$ can be dropped.

It is in particular the case when $g_{kt} \sim \nu_k$ $\forall t$ in an independent way

In this stochastic model:

$$\mathbb{E}[R_T] = \mathbb{E}\Big[\max_{k=1..K} \sum_{t=1}^{T} g_{kt}\Big] - \mathbb{E}\Big[\sum_{t=1}^{T} g_{I_t t}\Big]$$

$$\geq T \max_{k=1..K} \mathbb{E}[g_{k1}] - \mathbb{E}\Big[\sum_{t=1}^{T} Y_t\Big]$$

$$= T \max_{k=1..K} \mathbb{E}(\nu_k) - \mathbb{E}\Big[\sum_{t=1}^{T} Y_t\Big] = \overline{R}_T$$

($g_{I_t t}$ is Y_t)

The adversarial results entail in particular that there exists a strategy of the decision-maker such that

$$\sup_{\nu_1 \dots \nu_K \in \mathcal{P}([0,1])} \overline{R}_T \leq \sup_{\substack{\text{opponents} \\ \text{picking } g_{jt} \in [0,1]}} \mathbb{E}[R_T] \leq C\sqrt{TK}$$

for some numerical constant C

while the ranking of lower bounds is:

For all (randomized) strategies of the decision-maker, for all $K \geq 2$ and $T \geq K/5$,

$$\mathbb{E}[R_T] \geq \sup_{\substack{\text{oppnents} \\ \text{picking } g_{jt} \in [0,1]}} \mathbb{E}[R_T] \geq \sup_{\substack{\text{individual} \\ \text{sequences } g_{jt} \in [0,1]}} \mathbb{E}[R_T] \geq \sup_{\nu_1, \dots \nu_K \in \mathcal{P}([0,1])} \overline{R}_T \geq \frac{1}{20}\sqrt{TK}$$

and even: sup over $\nu_1 \dots \nu_K$ being Bernoulli distributions

as proved last week

Conclusion:

The minimax rates of the regret for stochastic bandits on $[0,1]$ or against oblivious opponents picking rewards in $[0,1]$ are $\sqrt{TK}$.

But: What is the minimax rate against general, reactive opponents?

- Should/Can the $\sqrt{TK \ln K}$ upper bound be improved?
- Should/Can the $\sup_{\text{opponents}} \mathbb{E}[R_T]$ lower bound be improved?

look for sequences of payoffs g_{kt} with real and strong correlations/dependencies in the past