

Pour aller plus loin sur le thème des “adversarial bandits” (à défaut d’avoir pu mieux voir ce sujet en cours; ce qui suit ne sera pas testé l’examen)

Links between:

- Distribution-free regret bounds for stochastic bandits
- Adversarial bandits

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

Stochastic bandits

With each arm a is associated $\vec{\mu}_a \in \mathcal{P}([q,1])$

For $t=1,2,\dots$

- The decision maker picks $I_t \in \{1..K\}$
- Her reward Y_t , which is such that $Y_t | I_t \sim \vec{\mu}_{I_t}$, is her only piece of information

Aim: control the regret

$$R_T = T \max_{a=1..K} E(\vec{\mu}_a) - E\left[\sum_{t=1}^T Y_t\right]$$

Adversarial bandits

An opponent selects the payoffs g_{I_t}

For $t=1,2,\dots$

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

Aim: control the regret

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

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 $[q,1]$, for all $T \geq 1$,

$$\begin{cases} \text{with probability at least } 1-\delta, \\ E[R_T] \leq C \sqrt{TK \ln(K/\delta)} \end{cases}$$

where the probability and E are w.r.t decision-maker's internal randomization

$$R_T \leq C \sqrt{TK \ln(K/\delta)}$$

for some numerical constant C

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

It is in particular the case when $g_{I_t} \sim \vec{\mu}_k$ i.e. in an independent way

In this stochastic model:

$$\begin{aligned}
 E[R_T] &= E\left[\max_{k=1..K} \sum_{t=1}^T g_{kt}\right] - E\left[\sum_{t=1}^T g_{I_t,t}\right] \quad \leftarrow g_{I_t,t} \text{ is } y_t \\
 &\geq T \max_{k=1..K} E[g_{k,1}] - E\left[\sum_{t=1}^T y_t\right] \\
 &= T \max_{k=1..K} E(\bar{y}_k) - E\left[\sum_{t=1}^T y_t\right] = \bar{R}_T
 \end{aligned}$$

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

$$\sup_{\vec{y}_1, \dots, \vec{y}_K \in \mathcal{P}([0,1])} \bar{R}_T \leq \sup_{\substack{\text{opponents} \\ \text{picking } g_{jt} \in [0,1]}} E[R_T] \leq C \sqrt{TK} \quad \text{for some numerical constant } C$$

while the ranking of lower bounds is:

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

$$\begin{aligned}
 \sup_{\substack{\text{opponents} \\ \text{picking } g_{jt} \in [0,1]}} E[R_T] &\geq \sup_{\substack{\text{individual} \\ \text{sequences } g_{jt} \in [0,1]}} E[R_T] \geq \sup_{\vec{y}_1, \dots, \vec{y}_K \in \mathcal{P}([0,1])} \bar{R}_T \geq \frac{1}{20} \sqrt{TK} \\
 &\quad \uparrow \text{and even:} \quad \uparrow \text{as proved last week} \\
 &\quad \text{sup over } \vec{y}_1, \dots, \vec{y}_K \text{ being Bernoulli distributions}
 \end{aligned}$$

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 generally reactive opponents?

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

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

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

$$\ell_t = (\ell_{jt})_{j=1,\dots,N} \text{ with } \ell_{jt} \in [0, M] \text{ and } I_t \sim p_t,$$

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

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

She resorts to the (conditionally) unbiased estimators

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

Denoting by $\mathcal{F}_{t-1} = \sigma(p_s, I_s, \ell_s, s \leq t-1)$ and possibly some auxiliary randomizations

$$\text{and } \mathcal{F}_0 = \{\emptyset, \Omega\}$$

we showed:

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

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:

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

↑ exploitation term
 ↑ exploration term

Hence the name Exp.3 for: exponential weights
for exploration and exploitation

Then: $\hat{l}_{jt} \in [0, \frac{MN}{\gamma_t}]$ as $p_{jt} \geq \gamma_t/N$ and $l_{jt} \in [0, M]$

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

$$(*) \quad \sum_{t=1}^T \hat{l}_{jt} \leq \sum_{t=1}^T \underbrace{E[\hat{l}_{jt} | \mathcal{F}_{t-1}]}_{= l_{jt}} + MN \underbrace{\sqrt{\sum_{t=1}^T \frac{1}{2\gamma_t^2} \ln \frac{1}{\delta}}}_{\text{depends on the range}}$$

A sharper bound is in terms of the conditional variances:

$$\text{Var}_{\mathcal{F}_{t-1}}(\hat{l}_{jt}) \leq E[\hat{l}_{jt}^2 | \mathcal{F}_{t-1}] \leq M^2 E[\mathbb{1}_{l_{jt} \neq 0} / p_{jt}^2 | \mathcal{F}_{t-1}]$$

$\uparrow$
 as $l_{jt}^2 \leq M^2$

$$= M^2 / p_{jt} \leq \frac{M^2 N}{\gamma_t}$$

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

$$(**) \quad \sum_{t=1}^T \hat{l}_{jt} \leq \sum_{t=1}^T l_{jt} + O\left(MN \sqrt{\sum_{t=1}^T \frac{1}{\gamma_t} \ln \frac{1}{\delta}}\right)$$

→ 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).

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

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 inconvenient 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:

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

or put differently, $\mathbb{E}[e^{\eta(X - \mathbb{E}[X | \mathcal{G}_j])} | \mathcal{G}_j] \leq \exp\left(\frac{e^{\eta M} - \eta M - 1}{\eta^2 M^2} \text{Var}_{\mathcal{G}_j}(X)\right)$ $\uparrow$ conditional variance of X

Proof: (sketch) $\eta(X - \mathbb{E}[X | \mathcal{G}_j]) \leq \eta M$ and $x \in \mathbb{R} \mapsto \frac{e^x - x - 1}{x^2}$ 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 $\mathbb{E}[\cdot | \mathcal{G}_j]$:
$$\mathbb{E}[e^{\eta(X - \mathbb{E}[X | \mathcal{G}_j])} | \mathcal{G}_j] - 1 \leq \text{Var}_{\mathcal{G}_j}(X) \frac{e^{\eta M} - \eta M - 1}{M^2}$$

as $\ln u \leq u - 1$ $\forall u > 0$ $\hookrightarrow \geq \ln \mathbb{E}[e^{\eta(X - \mathbb{E}[X | \mathcal{G}_j])} | \mathcal{G}_j]$, 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: $\forall \varepsilon > 0, \quad \forall V > 0,$

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

- Statistical version:

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

where V is a numerical constant > 0

$$\text{and } \sum_{t=1}^T \text{Var}_{\mathcal{F}_{t-1}}(X_t) \leq V$$

Proof: $\eta > 0$,

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

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

$$\text{Thus } \mathbb{P} \left\{ \underbrace{\sum_{t=1}^T (X_t - \mathbb{E}[X_t | \mathcal{F}_{t-1}])}_{= S_T} \geq \varepsilon \quad \text{and} \quad \underbrace{\sum_{t=1}^T \text{Var}_{\mathcal{F}_{t-1}}(X_t)}_{= \sigma_T^2} \leq V \right\}$$

$$\stackrel{\eta > 0}{=} \mathbb{P} \left\{ \underbrace{e^{\eta S_T} - \frac{(e^{\eta M} - \eta M - 1)}{\eta^2 M^2} \sigma_T^2}_{= M_T} \geq e^{\eta \varepsilon - \frac{(e^{\eta M} - \eta M - 1)}{\eta^2 M^2} \sigma_T^2} \quad \text{and} \quad \sigma_T^2 \leq V \right\}$$

$$\leq \mathbb{P} \left\{ M_T \geq \exp \left(\eta \varepsilon - \frac{(e^{\eta M} - \eta M - 1)}{\eta^2 M^2} V \right) \right\}$$

$$\stackrel{\text{Markov's inequality}}{\leq} \frac{\exp \left(-\eta \varepsilon + \frac{(e^{\eta M} - \eta M - 1)}{\eta^2 M^2} V \right)}{\underbrace{\mathbb{E}[M_T]}_{\leq \mathbb{E}[M_0] = 1 \text{ as } (M_t)_t \text{ is a supermartingale}}}$$

to be optimized over $\eta > 0$:

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

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

$$f''(x) = \frac{Ve^x}{M} > 0$$

unique minimizer x^* s.t.

$$f'(x^*) = 0, \text{ i.e. } e^{x^*} = 1 + \frac{M\varepsilon}{V}$$

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

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

The bound is

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

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

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

$$\text{where } h(u) = (1+u) \ln(1+u) - u$$

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

$$\begin{aligned} \frac{V}{M^2} h\left(\frac{M\varepsilon}{V}\right) &\geq \frac{V}{M^2} \left(\frac{M\varepsilon}{V}\right)^2 / \left(2 + \frac{2}{3} \frac{M\varepsilon}{V}\right) \\ &= \varepsilon^2 / \left(V \left(2 + \frac{2}{3} \frac{M\varepsilon}{V}\right)\right) \end{aligned}$$

↑
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 $\exp\left(-\varepsilon^2 / \left(2V + \frac{2}{3} M \varepsilon\right)\right) \leq \delta$.

Indeed,

$$\begin{aligned} \varepsilon^2 &= \left(\sqrt{2V \ln \frac{1}{\delta}} + \frac{2}{3} M \ln \frac{1}{\delta}\right)^2 && \text{where } \varepsilon \geq \sqrt{2V \ln \frac{1}{\delta}} \\ &\geq 2V \ln \frac{1}{\delta} + \frac{2}{3} M \varepsilon \ln \frac{1}{\delta} \\ &= \left(2V + \frac{2}{3} M \varepsilon\right) \ln \frac{1}{\delta}. \end{aligned}$$

Application: Performance bound for Exp3

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

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

Remarks: - Not quite the $\sqrt{T}$ rate we wanted! $\hookrightarrow$ 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 $\sum \hat{\ell}_{it} - \sum \ell_{it} \leq \dots \rightarrow$ to be improved.

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

(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

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

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

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

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

Therefore,

$$\sum_{t=1}^T \sum_{j=1}^N p_{jt} \hat{\ell}_{jt} - \min_{i=1 \dots N} \sum_{t=1}^T \hat{\ell}_{it}$$

We already saw that:

$$\sum_j p_{jt} \hat{\ell}_{jt} = \ell_{I_t} \quad \text{and} \quad \sum_j p_{jt} \hat{\ell}_{jt}^2 \leq M^2 \frac{1_{\{I_t=j\}}}{p_{jt}}$$

Bernstein's inequality applied $3N$ times:

$$\forall i, \sum_{t=1}^T \hat{L}_{it} \leq \sum_{t=1}^T L_{it} + M\sqrt{N} \sqrt{\sum_{t=1}^T \frac{2}{\gamma_t} \ln \frac{3N}{\delta}} + \frac{2}{3} \frac{MN}{\gamma_T} \ln \frac{3N}{\delta}$$

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

conditional variance $\leq \frac{1}{p_j} \leq \frac{N}{\gamma_t}$

$$\forall j, \sum_{t=1}^T \gamma_t \hat{L}_{jt} \leq \sum_{t=1}^T \gamma_t \underbrace{L_{jt}}_{\leq M} + \sqrt{2 \sum_{t=1}^T \gamma_t M^2 N \ln \frac{3N}{\delta}} + \frac{2}{3} MN \ln \frac{3N}{\delta}$$

conditional variance $\leq \gamma_t^2 M^2 N / \gamma_t = M^2 N \gamma_t$

↑ 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 $\geq 1-\delta$,

$$R_T = \sum_{t=1}^T L_{I_t,t} - \min_{i=1,\dots,N} \sum_{t=1}^T L_{it} \leq \frac{\ln N}{\eta_T} + M\sqrt{N} \sqrt{\sum_{t=1}^T \frac{2}{\gamma_t} \ln \frac{3N}{\delta}} + M \sum_{t=1}^T \gamma_t + \frac{M^2}{2} \sum_{t=1}^T \frac{\eta_t}{1-\gamma_t} \quad [+ \text{ MANY OTHER TERMS }]$$

but looking only at the γ_t we see the issue:

$$\gamma_t \sim t^{-\alpha} \quad \rightarrow \quad \left(\sum \frac{1}{\gamma_t} \right)^{1/2} = O(t^{-(\alpha+1)/2})$$

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

Choose α s.t. $-\alpha+1 = 2(\alpha+1)$ i.e. $\alpha = -1/3$

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)

$$\hat{\ell}_{jt} = M - \frac{M - \ell_{jt}}{p_{jt}} \mathbb{1}_{j \neq j^*} - \frac{\beta_t}{p_{jt}}$$

Algorithm:

$$p_{jt} = (1 - \gamma_t) \exp\left(-\eta_t \sum_{s=1}^{t-1} \hat{\ell}_{js}\right) / \sum_{k=1}^N \exp\left(-\eta_t \sum_{s=1}^{t-1} \hat{\ell}_{ks}\right) + \gamma_t$$

Prove that for well-chosen η_t , γ_t and β_t :

$$R_T \leq O\left(\sqrt{TN \ln \frac{N}{\delta}}\right) \quad \text{with probability at least } 1 - \delta.$$

NOTE: I'm sorry, I will NOT have time to write up corrections!