Optimality of the $\sqrt{\frac{T}{2}\ln N}$ bound: asymptotic lower bound.

In the case of linear losses

By homogeneity we may assume that $\ell_{jt} \in [a,1]$ (ie, $m=0$ and $M=1$) for all t and j.

Theorem: All forecasting strategies of the statistician are such that

$$1 \leqslant \liminf_{N\to+\infty} \liminf_{T\to+\infty} \sup_{\ell_{jt}\in[a,1]} \frac{\sum_{t,j} p_{jt}\ell_{jt} - \min_k \sum_{t=1}^T \ell_{kt}}{\sqrt{\frac{T}{2}\ln N}}$$

the opponent does not need to react, it suffices to consider fixed-in-advance sequences (individual sequences)

Proof: We lower bound the $\sup_{\ell_{jt}}$ by an expectation $\mathbb{E}$, assuming that the losses are given by random variables

L_{jt} iid $\sim$ Ber$(1/2)$

$\mathcal{F}_{t-1} = \sigma(L_{js},\ s\leqslant t-1 \text{ and } j\in\{1,\dots N\})$ is the information available at the beginning of round t: thus, p_t is $\mathcal{F}_{t-1}$-measurable

and by the tower rule,

$$\mathbb{E}\Big[\sum_j p_{jt}L_{jt}\Big] = \mathbb{E}\Big[\sum_j p_{jt}\,\underbrace{\mathbb{E}[L_{jt}\mid\mathcal{F}_{t-1}]}_{=\mathbb{E}[L_{jt}]=1/2}\Big] = 1/2$$

by independence: $p_{jt} \perp\!\!\!\perp L_{jt}$

$\mathbb{E}[p_{jt}L_{jt}] = \mathbb{E}[p_{jt}]\,\mathbb{E}[L_{jt}] = \mathbb{E}[p_{jt}]\,1/2$

$\sum_j \mathbb{E}[p_{jt}L_{jt}] = 1/2$

not so surprising: you cannot predict iid random variables!

* But *

we will show that

$\mathbb{E}\Big[\min_k \sum_{t=1}^T L_{kt}\Big]$ is smaller than $T/2$, even if "individually", we have $\forall k,\ \mathbb{E}\Big[\sum_{t=1}^T L_{kt}\Big] = T/2$

"crowd" effect

(due to the central limit theorem)

Let us summarize what we have so far:

worst-case regret $= \sup_{\ell_{jt} \in [0,1]} \left\{ \sum_{jt} p_{jt} \ell_{jt} - \min_k \sum_{t=1}^T \ell_{kt} \right\} \geq \mathbb{E}\left[\sum_{jt} p_{jt} L_{jt} - \min_k \sum_{t=1}^T L_{kt} \right]$

$= \frac{T}{2} - \mathbb{E}\left[\min_k \sum_{t=1}^T L_{kt} \right]$

$= \frac{\sqrt{T}}{2} \mathbb{E}\left[\max_{k=1..N} \frac{\sum_{t=1}^T \left(\frac{1}{2} - L_{kt}\right)}{\frac{1}{2}\sqrt{T}} \right]$

we have that $\mathbb{E}[Z_{kT}] = 0$ and $\mathbb{V}(Z_{kT}) = 1$, by independence we denote $Z_{kT} = \frac{\sum_{t=1}^T \left(\frac{1}{2} - L_{kt}\right)}{\frac{1}{2}\sqrt{T}}$

By the central limit theorem, $Z_{kT} \xrightarrow[T \to +\infty]{\mathcal{L}} \mathcal{N}(0,1)$

We will show successively that

(1) $\mathbb{E}\left[\max_{k=1..N} Z_{kT} \right] \xrightarrow[T \to +\infty]{} \mathbb{E}\left[\max_{k \leq N} G_k \right]$ where $G_1, \dots, G_k$ iid $\sim \mathcal{N}(0,1)$

would be sufficient to prove $\liminf_{T \to +\infty} \mathbb{E}[\max_k Z_{kT}] \geq \mathbb{E}[\max_k G_k]$

(2) $\mathbb{E}\left[\max_{k \leq N} G_k \right] \underset{N \to +\infty}{\sim} \sqrt{2 \ln N}$

would be sufficient to prove $\liminf_{T \to +\infty} \frac{\mathbb{E}[\max_{k \leq N} G_k]}{\sqrt{2 \ln N}} \geq 1$

which will conclude the proof of the theorem.

Point (1) relies on asymptotic uniform integrability.

Note: in general, $Y_T \xrightarrow{\mathcal{L}} Y$ does NOT imply $f(Y_T) \to f(Y)$ if f is just assumed continuous.
Ex: $Y_T = T^2 \mathbb{1}_{U_T \in [0, 1/T]}$ where U_t iid $\sim \mathcal{U}[0,1]$
We have $Y_T \to 0$ in probability but $\mathbb{E}Y_T = T \to +\infty$

Reminder: If $(Y_T)_{T \geq 1}$ is a sequence of $\mathbb{R}^N$-valued random variables with $Y_T \xrightarrow[T \to +\infty]{\mathcal{L}} Y$, then for all continuous and bounded functions $f: \mathbb{R}^N \to \mathbb{R}$,

$\mathbb{E}[f(Y_T)] \to \mathbb{E}[f(Y)]$

Here: $Y_T = (Z_{kT})_{k \in \{1,\dots,N\}} \in \mathbb{R}^N$

and $Y_T \xrightarrow{\mathcal{L}} \begin{bmatrix} G_1 \\ \vdots \\ G_N \end{bmatrix} \sim \mathcal{N}(0, I_N)$ by independence of the components

* but * $f(z_1 \dots z_N) = \max_{k \le N} z_k$ is not bounded (though being continuous).

Definition: $(f(Y_T))_{T \ge 1}$ is a.u.i. (asymptotically uniformly integrable) if

$$\lim_{L \to +\infty} \limsup_{T \to +\infty} \mathbb{E}\left[|f(Y_T)| \, \mathbb{1}_{\{|f(Y_T)| > L\}} \right] = 0$$

Lemma: If $\left\{ \begin{array}{l} Y_T \xrightarrow{\mathcal{L}} Y \\ (f(Y_T)) \text{ is u.a.i} \\ f \text{ continuous} \end{array} \right\}$ then (a) $f(Y) \in \mathbb{L}^1$ and $f(Y_T) \in \mathbb{L}^1$ $\forall T$, and (b) $\mathbb{E}[f(Y_T)] \xrightarrow[T \to +\infty]{} \mathbb{E}[f(Y)]$

of (a): see an added page later in this document

Proof: $\hookrightarrow$ (b) We may assume with no loss of generality that $f \ge 0$

(otherwise, we consider $f^+ = \max\{f, 0\}$ and $f^- = \max\{-f, 0\}$ the positive and negative parts of f; and use $f = f^+ - f^-$ and the fact that $f(Y) \in \mathbb{L}^1$)

$\forall L$:

$$\begin{aligned} & \left| \mathbb{E}[f(Y)] - \mathbb{E}[f(Y_T)] \right| \\ \le & \left| \mathbb{E}[f(Y)] - \mathbb{E}[\min\{f(Y), L\}] \right| \\ & + \left| \mathbb{E}[\min\{f(Y), L\}] - \mathbb{E}[\min\{f(Y_T), L\}] \right| \\ & + \left| \mathbb{E}[\min\{f(Y_T), L\}] - \mathbb{E}[f(Y_T)] \right| \end{aligned}$$

$\rightsquigarrow$ tends to 0 as $T \to +\infty$, since $\min\{f(\cdot), L\}$ is continuous, bounded in $[0, L]$; and $Y_T \xrightarrow{\mathcal{L}} Y$

Thus, $\forall L$:

$$\begin{aligned} & \limsup_{T \to +\infty} \left| \mathbb{E}[f(Y)] - \mathbb{E}[f(Y_T)] \right| \\ \le & \left| \mathbb{E}[f(Y)] - \mathbb{E}[\min\{f(Y), L\}] \right| \\ & + \limsup_{T \to +\infty} \left| \mathbb{E}[\min\{f(Y_T), L\}] - \mathbb{E}[f(Y_T)] \right| \end{aligned}$$

Taking $\limsup_{L \to +\infty}$ in both sides (the left-hand side is independent of L)

and using that $\limsup_{T \to +\infty} (a_T + b_T) \le \limsup_{T \to +\infty} a_T + \limsup_{T \to +\infty} b_T$

[think in terms of largest adherence values]

we get

$$\begin{aligned} \limsup_{T \to +\infty} \left| \mathbb{E}[f(Y)] - \mathbb{E}[f(Y_T)] \right| \le & \limsup_{L \to +\infty} \left| \mathbb{E}[f(Y)] - \mathbb{E}[\min\{f(Y), L\}] \right| \\ & + \limsup_{L} \limsup_{T} \left| \mathbb{E}[\min\{f(Y_T), L\}] - \mathbb{E}[f(Y)] \right| \end{aligned}$$

By monotone convergence, (or dominated)

$$\mathbb{E}\left[\min\{f(Y), L\}\right] \xrightarrow[L\to+\infty]{} \mathbb{E}[f(Y)]$$

And by the u.a.i property,

$$\limsup_{L\to+\infty} \limsup_{T\to+\infty} \left| \mathbb{E}\left[\min\{f(Y_T), L\}\right] - \mathbb{E}[f(Y_T)] \right| = 0$$

$$= \left| \mathbb{E}\left[(f(Y_T) - L)\, \mathbb{1}_{\{f(Y_T)>L\}} \right] \right|$$

because f assumed ≥ 0 ↳ $\leq \mathbb{E}\left[f(Y_T)\, \mathbb{1}_{\{f(Y_T)>L\}} \right]$

Application: Need only to check that $\max_{k\leq N} Z_{kT}$ is u.a.i to get (1) from the lemma.

Indeed, given $L>0$,

$$\mathbb{E}\left[\left|\max_{k\leq N} Z_{kT}\right| \mathbb{1}_{\left\{\left|\max_{k\leq N} Z_{kT}\right|>L\right\}} \right]$$

$$= \int_0^{+\infty} \mathbb{P}\left(\left|\max_{k\leq N} Z_{kT}\right| \mathbb{1}_{\left\{\left|\max_{k\leq N} Z_{kT}\right|>L\right\}} \geq u \right) du$$

a well-known fact: if $X\geq 0$,

$$\mathbb{E}[X] = \int_0^{+\infty} \mathbb{P}(X\geq u)\, du$$

Indeed, by Fubini-Tonelli:

$$\mathbb{E}[X] = \int_{[0,+\infty)} x\, d\mathbb{P}^X(x) = \int_{[0,+\infty)^2} \mathbb{1}_{\{0\leq u\leq x\}}\, d\mathbb{P}^X(x)\, du$$

$$= \int_{[0,+\infty)} \mathbb{P}(X\geq u)\, du$$

equals

$$\begin{cases} \mathbb{P}\left(\left|\max_{k\leq N} Z_{kT}\right| > L \right) & \text{if } u \leq L \\ \mathbb{P}\left(\left|\max_{k\leq N} Z_{kT}\right| \geq u \right) & \text{if } u > L \end{cases}$$

But, for all $u>0$:

$$\mathbb{P}\left(\left|\max_{k\leq N} Z_{kT}\right| \geq u \right) \leq \mathbb{P}\left(\exists j\leq N \mid |Z_{jT}|\geq u \right)$$ ↳ union bound + same distribution

$$\leq N \times \mathbb{P}(|Z_{1T}|\geq u)$$

$$\leq N \times \frac{\operatorname{Var} Z_{1T}}{u^2} = N/u^2$$ ↳ $\mathbb{E}Z_{1T}=0$ and Tchebychev

cf. $\operatorname{Var} Z_{1T}=1$

Substituting above, the conclusion is obtained:

$$\mathbb{E}\left[\left|\max_{k\leq N} Z_{kT}\right| \mathbb{1}_{\left\{\left|\max_{k\leq N} Z_{kT}\right|>L\right\}} \right] \leq \int_0^L \frac{N}{L^2}\, du + \int_L^{+\infty} \frac{N}{u^2}\, du = \frac{2N}{L}$$

bound independent of T and such that $\to 0$ as $L\to+\infty$

Note (credits: Clémençon Besserfield + Chenlin Gu.)

follows from $Y_T \to Y$ & f continuous

Proof of property (a): $(f(Y_T))$ u.a.i and $f(Y_T) \to f(Y)$ implies $f(Y) \in \mathbb{L}^1$ (and $f(Y_T) \in \mathbb{L}^1$ $\forall T$)

Indeed: By uai, $\exists L_0$ | $\limsup_{T \to +\infty} \mathbb{E}\left[|f(Y_T)| \mathbb{1}_{\{|f(Y_T)| > L_0\}} \right] < +\infty$

in particular, $\limsup_{T \to +\infty} \mathbb{E}\left[|f(Y_T)| \right] < +\infty$

Thus, $\left(\mathbb{E}\left[|f(Y_T)| \right] \right)_{T \geq 1}$ is bounded

By Skorokhod, we can assume with no loss of generality that

$f(Y_T) \to f(Y)$ a.s.

so that Fatou's lemma indicates $\mathbb{E}\left[|f(Y)| \right] \leq \liminf_{T \to +\infty} \mathbb{E}\left[|f(Y_T)| \right] < +\infty$

It only remains to prove (2): $\mathbb{E}\left[\max_{k\leq N} G_k\right] \sim \sqrt{2\ln N}$

where G_j iid $\sim \mathcal{N}(0,1)$

* Easy (and useless part for our proof): $\mathbb{E}\left[\max_{k\leq N} G_k\right] \leq \sqrt{2\ln N}$

<u>Lemma</u>: If $V_1, \dots V_N$ are such that $\mathbb{E}[e^{\lambda V_j}] \leq e^{\lambda^2\sigma^2/2}$ $\forall j$ $\forall \lambda > 0$

then $\mathbb{E}\left[\max_{k\leq N} V_k\right] \leq \sqrt{2\sigma^2 \ln N}$

<u>Note</u>: Valid even without independence → independence will be needed for the lower bound

If $G_k \sim \mathcal{N}(0,\sigma^2)$ then $\mathbb{E}[e^{\lambda G_k}] = e^{\lambda^2\sigma^2/2}$

If $G_k \in [m,M]$ with $\mathbb{E}[G_k] = 0$ then (Hoeffding's lemma) $\mathbb{E}[e^{\lambda G_k}] \leq e^{\lambda^2 (M-m)^2/8}$

↳ We call subgaussian the random variables V_k satisfying the assumptions of the lemma.

<u>Proof</u>: Pisier's argument:

$$\mathbb{E}\left[\max_{k\leq N} V_k\right] \leq \frac{1}{\lambda} \ln \mathbb{E}\left[e^{\lambda \max_{k\leq N} V_k}\right] \quad \text{(Jensen's ineq. } \forall \lambda > 0\text{)}$$

$$\leq \frac{1}{\lambda} \ln \mathbb{E}\left[\sum_{k=1}^{N} e^{\lambda V_k}\right] \leq \frac{1}{\lambda} \ln\left(N e^{\lambda^2\sigma^2/2}\right)$$

$$= \frac{\ln N}{\lambda} + \lambda \frac{\sigma^2}{2} = \sqrt{2\sigma^2 \ln N}$$

choose $\lambda^* = \sqrt{\frac{2\ln N}{\sigma^2}}$

* « Difficult » part of the proof: $\liminf_{N\to+\infty} \frac{\mathbb{E}\left[\max_{k\leq N} G_k\right]}{\sqrt{2\ln N}} \geq 1$

(Credit: Pascal Massart explained to me how to proceed efficiently!)

We use: let $F_N = \Phi^N$ be the cumulative distribution function of $\max_{k\leq N} G_k$, where Φ is the cumulative distribution function of the $\mathcal{N}(0,1)$ distribution.

F_N is a bijection $\mathbb{R} \to (0,1)$ so that, for $U \sim \mathcal{U}_{[0,1]}$,

$$F_N^{-1}(U) = \Phi^{-1}(U^{1/N}) \quad \text{and} \quad \max_{k \le N} G_k \quad \text{have the}$$

same distribution.

In particular,

$$\mathbb{E}\left[\max_{k \le N} G_k\right] = \mathbb{E}\left[\Phi^{-1}(U^{1/N})\right] = \int_0^1 \Phi^{-1}(u^{1/N})\, du$$

we separate the integrals depending on the sign of the integrand.

$$= \underbrace{\int_0^{1/2} \underbrace{\Phi^{-1}(u^{1/N})}_{\le 0}\, du}_{\le 0} + \underbrace{\int_{1/2}^1 \underbrace{\Phi^{-1}(u^{1/N})}_{\ge 0}\, du}_{\ge 0}$$

$$\mathbb{E}\left[\max_{k \le N} G_k \, \mathbb{1}_{\{\max_{k \le N} G_k \le 0\}}\right]$$

is ≤ 0

and, is

$$\ge \mathbb{E}\left[G_1 \, \mathbb{1}_{\{\max_{k \le N} G_k \le 0\}}\right]$$

$$= \mathbb{E}\left[G_1 \, \mathbb{1}_{\{G_1 \le 0\}} \, \mathbb{1}_{\{\max_{k \le N} G_k \le 0\}}\right]$$

$$\ge \mathbb{E}\left[G_1 \, \mathbb{1}_{\{G_1 \le 0\}}\right]$$

so

$$\left| \mathbb{E}\left[\max_{k \le N} G_k \, \mathbb{1}_{\{\max_{k \le N} G_k \le 0\}}\right] \right|$$

$$\le \left| \mathbb{E}\left[G_1 \, \mathbb{1}_{\{G_1 \le 0\}}\right] \right|$$

$$\le \sqrt{\mathbb{E}G_1^2} \sqrt{\mathbb{P}\{G_1 \le 0\}} = 1/\sqrt{2}$$

Cauchy Schwarz

fix any $\delta > 0$; for N large enough, $\delta \ge 1/2^N$

$$\ge \int_\delta^1 \Phi^{-1}(u^{1/N})\, du \ge (1-\delta)\, \Phi^{-1}(\delta^{1/N})$$

Φ^{-1} increasing

Thus we have proved so far:

$$\liminf_{N \to +\infty} \frac{\mathbb{E}\left[\max_{k \le N} G_k\right]}{\sqrt{2 \ln N}} \ge \sup_{\delta \in (0,1)} \liminf_{N \to +\infty} \frac{(1-\delta)\, \Phi^{-1}(\delta^{1/N})}{\sqrt{2 \ln N}}$$

It's only a matter of real analysis / calculus now:

We will successfully prove

(a) $1-\Phi(x) \underset{x\to+\infty}{\sim} e^{-x^2/2} \Big/ x\sqrt{2\pi}$

(b) $\ln \frac{1}{1-v} \sim \left(\Phi^{-1}(v)\right)^2/2$ as $v \to 1$

(c) thus the desired result.

Proof of (a): by two integration by parts (IbP)

$$1-\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} e^{-t^2/2} dt = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} t e^{-t^2/2} \cdot \frac{1}{t} dt$$

$$\underset{\text{IbP}}{=} \underbrace{\left[-\frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t} \right]_x^{+\infty}}_{= \frac{e^{-x^2/2}}{x\sqrt{2\pi}}} - \underbrace{\frac{1}{\sqrt{2\pi}} \int_x^{+\infty} \frac{e^{-t^2/2}}{t^2} dt}$$

Second IbP:

$$0 \leq \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} t e^{-t^2/2} \frac{1}{t^3} dt$$

$$= \left[-\frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t^3} \right]_x^{+\infty} - \underbrace{\int_x^{+\infty}}_{\geq 0}$$

$$\text{thus} \leq \frac{1}{\sqrt{2\pi}} \frac{e^{-x^2/2}}{x^3}$$

so that the $e^{-x^2/2}/x\sqrt{2\pi}$ term was dominating in what the first IbP yielded, hence the result.

Proof of (b): Recall that if $f(x) \sim g(x)$ and $\ln f(x)$ is bounded away from 0 in the limit, then $\ln f(x) \sim \ln g(x)$

indeed: $\frac{\ln g(x)}{\ln f(x)} = 1 + \frac{\ln\left(g(x)/f(x)\right)}{\ln f(x)} \longrightarrow 1$

In particular

$$\ln \frac{1}{1-\Phi(x)} \underset{x\to+\infty}{\sim} \ln\left(e^{x^2/2} x\sqrt{2\pi}\right) \underset{x\to+\infty}{\sim} \frac{x^2}{2}$$

With $x = \Phi^{-1}(v) \longrightarrow +\infty$ as $v \to 1$ (by composing limits):

$$\ln \frac{1}{1-v} \underset{v\to 1}{\sim} \left(\Phi^{-1}(v)\right)^2/2$$

(c) $\Phi^{-1}(\delta^{1/N}) \sim \sqrt{2 \ln \frac{1}{1-\delta^{1/N}}}$ given $\delta^{1/N} \xrightarrow[\text{as } N\to+\infty]{<} 1$

$\sim \sqrt{2\ln N}$ as $N \to +\infty$

given $1-\delta^{1/N} = 1 - \exp(\frac{1}{N}\ln\delta) \sim -\frac{1}{N}\ln\delta = \frac{1}{N}\ln\frac{1}{\delta}$

thus:

$$\liminf_N \frac{\mathbb{E}[\max_{k\le N} G_k]}{\sqrt{2\ln N}} \geqslant \sup_{\delta\in(0,1)} \left\{ (1-\delta) \times \underbrace{\liminf_{N\to+\infty} \frac{\Phi^{-1}(\delta^{1/N})}{\sqrt{2\ln N}}}_{=1} \right\} = 1$$

Exercice — **Finite-time lower bound.** (also known as: Non-asymptotic lower bound)

It's a difficult exercise, try to do as much as possible! Part of the difficulty comes from the manipulation of Kullback-Leibler divergences, we will study their properties in detail in a few weeks, when studying the lower bounds for stochastic bandits.

Let $\mathbb{P}_0$ be a probability such that the losses ℓ_{kt} are $\forall k,t$ id $\sim \mathcal{B}(1/2)$

For $j \in \{1,\dots,N\}$ let $\mathbb{P}_j$ be a probability such that:

- all losses ℓ_{kt} are independent
- $\ell_{kt} \sim \mathrm{Ber}(1/2)$ $\forall t$, $\forall k \neq j$
- $\ell_{jt} \sim \mathrm{Ber}(1/2 - \varepsilon)$ $\forall t$

(1) Show that

$$\sup_{\ell_{jt} \in [0,1]} \left\{ \sum_{jt} p_{jt} \ell_{jt} - \min_k \sum_{t=1}^T \ell_{kt} \right\} \geq T\varepsilon \max_{k=1,\dots,N} \left\{ 1 - \frac{1}{T} \sum_{t=1}^T \mathbb{E}_k[p_{kt}] \right\}$$

REMINDER or QUICK INTRODUCTION TO A NEW OBJECT:

Recall that the Kullback-Leibler divergence between two probability distributions μ and ν equals:

$$KL(\nu,\mu) = \begin{cases} +\infty & \text{if } \mu \text{ not absolutely continuous wrt } \nu \\ \int \left(\frac{d\nu}{d\mu} \ln \frac{d\nu}{d\mu} \right) d\mu \in [0,+\infty] & \\ = \int \left(\ln \frac{d\nu}{d\mu} \right) d\nu & \text{if } \nu \ll \mu \end{cases}$$

(KL well-defined as $x \mapsto x \ln x$ is $\geq -1/e$ over $(0,+\infty)$)

KL has the following properties:

- $KL(\nu \otimes \nu', \mu \otimes \mu') = KL(\nu,\mu) + KL(\nu',\mu')$ ⇝ by Fubini-Tonelli
- $\forall$ random variables X, $KL(\nu^X, \mu^X) \leq KL(\nu,\mu)$ ⇝ data compression inequality (we take it for granted and will prove it later)

We denote by $kl(p,q) = KL(\mathrm{Ber}(p), \mathrm{Ber}(q))$

$$= p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}$$

the KL-divergence between two Bernoulli distributions.

(2) Prove Fano's lemma as revisited by Lucien Birgé:

> Let $(\Omega, \mathcal{F})$ be a measurable space
> Let $Q_1, \dots Q_N$ be probability distributions over Ω
> Let $(A_1 \dots A_N)$ be a partition of Ω
> then
> $$\min_{j=1\dots N} Q_j(A_j) \leq \max\left\{\frac{2e}{2e+1}, \frac{\bar{K}}{\ln N}\right\}$$
> where $\bar{K} = \frac{1}{N-1}\sum_{j=2}^{N} KL(Q_j, Q_1)$

Hints: Consider $p = \frac{1}{N-1}\sum_{j=2}^{N} Q_j(A_j)$ and $q = \frac{1}{N-1}\sum_{j=2}^{N} Q_1(A_j)$

and denote $a = \min_{j=1\dots N} Q_j(A_j)$. Show that:

- $kl(p,q) \leq \bar{K}$ / to that end, first prove that KL is jointly convex (it is a consequence of the data-processing inequality, explain why)
- $kl(p,q)/\ln N \geqslant a$ for $a \geqslant 2e/2e+1$

$KL(P,Q) \leqslant KL(E_1 \dots E_N, Q_1 \dots Q_N)$ for any random variable Z taking values in $(E_1 \dots E_N)$ equipped with $\mathcal{F}$ or Q

(3) Extend Fano's lemma to random variables so as to show

$$\min_{k=1,\dots N} \mathbb{E}_k\left[\frac{1}{T}\sum_{t=1}^{T} r_{kt}\right] \leqslant \max\left\{\frac{2e}{2e+1}, \frac{\bar{K}'}{\ln N}\right\}$$

where $\bar{K}' = \frac{1}{N-1}\sum_{j=2}^{N} KL(\mathbb{P}_j^L, \mathbb{P}_1^L)$ with $L = (L_{kt})_{\substack{t\leq T \\ k\leq N}}$.

(4) Show that $\bar{K}' \leq 5T\varepsilon^2$ for $\varepsilon \leq 1/10$.

(5) Conclude to the following theorem:

Theorem: For all strategies, for all $N \geq 2$ and $T \geqslant 17 \ln N$,

$$\sup R_T \geqslant 0.06\sqrt{T \ln N}$$

Convex functions & comparison to the best convex vector

↳ efficient forecaster (but worse bound...)

$\mathcal{X} = \{(p_1 \dots p_N) : \sum_j p_j = 1, \ \forall k \ p_k \geq 0\}$ is the simplex of convex weight vectors.

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

1. The statistician and the opponent pick simultaneously $p_t \in \mathcal{X}$ and a convex loss function $\ell_t : \mathcal{X} \to [m, M]$
2. ℓ_t and p_t are revealed

↳ **Regret**: $R_T = \sum_{t=1}^{T} \ell_t(p_t) - \inf_{p \in \mathcal{X}} \sum_{t=1}^{T} \ell_t(p)$ to be controlled in a uniform way

Fact: Convex functions are subdifferentiable on the interior of their domain of definition:

Let $f : \mathcal{D} \to \mathbb{R}$ be convex, where $\mathcal{D} \subseteq \mathbb{R}^N$ is convex:

$\forall x \in \mathring{\mathcal{D}}, \ \exists \partial f_x \in \mathbb{R}^N \mid \forall y \in \mathcal{D}, \ f(x) - f(y) \leq \partial f_x \cdot (x - y)$

Ex: $f(x) = |x|$, $\partial_j(x) = \begin{cases} \{-1\} & \text{if } x < 0 \\ \{+1\} & \text{if } x > 0 \\ [-1, 1] & \text{if } x = 0 \end{cases}$

$\partial_j(x) = \{$ the set of all possible such $\partial f_x\}$ is called the subgradient of f at x

If f is differentiable at x, then $\partial_j(x) = \{\nabla f(x)\}$.

Application: If $p_t \in \mathring{\mathcal{X}}$ (i.e. $p_{jt} > 0 \ \forall j$) then $\exists \partial \ell_t(p_t) \in \mathbb{R}^N \mid \forall p \in \mathcal{X}, \ \ell_t(p_t) - \ell_t(p) \leq \partial \ell_t(p_t) \cdot (p_t - p)$

Example: Meta-statistical framework: $\ell_t(p) = \left(\sum_j p_j f_{jt} - y_t\right)^2$

ℓ_t is differentiable, $\nabla \ell_t(p) = \left(2\left(\sum_j p_j f_{jt} - y_t\right) f_{jt}\right)_j$.

↳ Gradients in $[-G, G]^N$ with $G = 2B^2$ if $f_{jt}, y_t \in [0, B]$

Strategy: Exponentiated Gradients (EG) with learning rate $\eta > 0$

$p_1 = \left(\frac{1}{N}, \dots \frac{1}{N}\right)$ and

$$p_{jt} = \exp\Big(-\eta \sum_{s=1}^{t-1} (\partial \ell_s(p_s))_j\Big) \Big/ \sum_{k=1}^{N} \exp\Big(-\eta \sum_{s=1}^{t-1} (\partial \ell_s(p_s))_k\Big)$$

→ strategy easy to implement as soon as subgradients can be easily computed.

Theorem: Assume that the ℓ_t are picked such that

$$\forall t, \forall j, \forall p \in \mathcal{X}, \quad (\partial \ell_t(p))_j \in [-G, G]$$

Then the regret of EG tuned with $\eta > 0$ is controled as

$$R_T = \sum_{t=1}^{T} \ell_t(p_t) - \inf_{p \in \mathcal{X}} \sum_{t=1}^{T} \ell_t(p) \leq \frac{\ln N}{\eta} + \eta \frac{G^2}{2} T$$

In particular, the choice $\eta = \frac{1}{G}\sqrt{\frac{2 \ln N}{T}}$ leads to

$$R_T \leq G\sqrt{2T \ln N}$$

Proof: Fix a p. Since by construction (cf. exponential weights), $p_t \in \mathcal{X}$ we have

$$\ell_t(p_t) - \ell_t(p) \leq \partial \ell_t(p_t) \cdot (p_t - p)$$

Thus

$$R_T \leq \sup_{p \in \mathcal{X}} \sum_{t=1}^{T} \partial \ell_t(p_t) \cdot (p_t - p)$$

$$= \sup_{p \in \mathcal{X}} \sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \tilde{\ell}_{jt} - \sum_{t=1}^{T} \sum_{j=1}^{N} p_j \tilde{\ell}_{jt}$$

where we denoted by $\tilde{\ell}_{jt} = (\partial \ell_t(p_t))_j$ the components of the subgradients: we interpret them as pseudo-losses

By linearity in p of the last upper bound,

$$R_T \leq \sup_{p \in \mathcal{X}} \sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \tilde{\ell}_{jt} - \sum_{t=1}^{T} \sum_{j} p_j \tilde{\ell}_{jt} = \sum_{tj} p_{jt} \tilde{\ell}_{jt} - \min_k \sum_{t=1}^{T} \tilde{\ell}_{kt}$$

↳ We have performed a reduction to the linear case with vectors of losses $(\tilde{\ell}_{1t} \dots \tilde{\ell}_{Nt})$ where $\tilde{\ell}_{jt} \in [-G, G]$

hence the stated bound.

Application: Oracle inequalities → From individual sequences to stochastic sequences

Question: Let $(Y_1 \dots Y_T)$ be a sequence of iid random variables taking values in an arbitrary set $\mathcal{Y}$; and $Q : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a bounded loss function, convex in its first argument, with bounded subgradients:

$$G = \sup_{y \in \mathcal{Y}} \sup_{p \in \mathcal{X}^*} \| \nabla Q(p, y) \|_\infty < +\infty$$

Aim: Construct $\hat{\theta}_T = \hat{\theta}_T(Y_1 \dots Y_T)$ such that

$$\mathbb{E}[Q(\hat{\theta}_n, Y)] \leq \inf_{\theta \in \mathcal{X}} \mathbb{E}[Q(\theta, Y)] + \varepsilon_n$$

where Y is independent of the Y_t with the same distribution and $\varepsilon_n \searrow 0$.

(The expectation $\mathbb{E}$ is w.r.t. $Y_1 \dots Y_T$ and Y.)

Typical machine learning method: Empirical risk minimization, possibly in a regularized way:

$$\hat{\theta}_T \in \operatorname*{argmin}_{\theta \in \mathcal{X}} \left\{ \frac{1}{T} \sum_{t=1}^{T} Q(\theta, Y_t) + \lambda \, \mathrm{reg}(\theta) \right\}$$

with $\mathrm{reg}(\theta) = \|\theta\|_2$ or $\|\theta_1\|$ or ...

and λ to be tuned (called the regularization factor)

Our method: (as in our proof of Sion's lemma)

1) Pretend data is sequential, while of course it is batch

$\tilde{\theta}_1 = (1/N, \dots 1/N)$ and by induction, with $\eta = \frac{1}{G}\sqrt{\frac{2 \ln N}{T}}$:

$$\tilde{\theta}_{jt} = \exp\left(-\eta \sum_{s=1}^{t-1} \left(\nabla Q(\tilde{\theta}_s, Y_s)\right)_j\right) \Big/ \sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \left(\nabla Q(\tilde{\theta}_s, Y_s)\right)_k\right)$$

2) Consider an average: $\hat{\theta}_T = \frac{1}{T}\sum_{t=1}^{T} \tilde{\theta}_t$

(note: we do not use Y_1!)

Guarantee: $\varepsilon_T = G\sqrt{\frac{2\ln N}{T}}$

Proof: By the theorem on EG (exponentiated gradient) and given our choice of η:

$$\sum_{t=1}^{T} Q(\tilde{\theta}_t, Y_t) - \inf_{\theta\in\Theta} \sum_{t=1}^{T} Q(\theta, Y_t) \leq G\sqrt{2T\ln N}$$

Thus: $\forall \theta\in\Theta$, $\frac{1}{T}\sum_{t=1}^{T} Q(\tilde{\theta}_t, Y_t) \leq \frac{1}{T}\sum_{t=1}^{T} Q(\theta, Y_t) + \varepsilon_T$ with $\varepsilon_T = G\sqrt{\frac{2\ln N}{T}}$

Now, since $Y_1 \dots Y_T$ and Y are iid:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q(\theta, Y_t)\right] = \mathbb{E}[Q(\theta, Y)]$$

We conclude the proof by showing that $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q(\tilde{\theta}_t, Y_t)\right] \geq \mathbb{E}[Q(\hat{\theta}_T, Y)]$

Indeed: $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q(\tilde{\theta}_t, Y_t)\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q(\tilde{\theta}_t, Y)\right]$

$\tilde{\theta}_t$ only depends on $Y_1 \dots Y_{t-1}$ and $(Y_1 \dots Y_{t-1}, Y_t)$ and $(Y_1 \dots Y_{t-1}, Y)$ have the same distribution

$\geq Q\left(\frac{1}{T}\sum_t \tilde{\theta}_t, Y\right)$

by convexity and because the second argument is now the same for everyone!

Extension: Oracle inequality for stationary data.

A stationary sequence $(Y_1, Y_2, \dots)$ is by definition a sequence of random variables such that

$\forall k \geq 1$, $\forall t \geq 1$, $(Y_1 \dots Y_k)$ and $(Y_{t+1}, \dots Y_{t+k})$ have the same distribution.

Construct $\hat{\theta}_T = \hat{\theta}_T(Y_1, \dots Y_T)$ such that, under the same conditions on Q as above:

$$\mathbb{E}[Q(\hat{\theta}_T, Y_{T+1})] \leq \inf_{\theta} \mathbb{E}[Q(\theta, Y_{T+1})] + G\sqrt{\frac{2\ln N}{T}}.$$

Hint: the sequential ordering of the Y_t may be different from the "natural" order used in the iid case. Try to mimic the key equality in distribution used in the iid case, that is, the fact that $(\tilde{\theta}_t, Y_t)$ and $(\tilde{\theta}_t, Y_{T+1})$ have the same distribution for some well-chosen $\tilde{\theta}_t$.

Solution: see next pages; not very clear (I hope that I ended up being clearer than that in class...).

Exercise / Oracle inequality for stationary data.

For a sequence of observations $y_1, \dots y_t \in \mathcal{Y}$, EG outputs the following weights, which we re-define as functions of the information available:

$$\psi_1^{EG} = \left(\tfrac{1}{N}, \dots \tfrac{1}{N}\right)$$

and for $t \geq 2$:

$$\psi_{jt}^{EG}(y_1, \dots y_{t-1}) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \left(\partial Q\left(\psi_s^{EG}(y_1 \dots y_{s-1}), y_s\right)\right)_j\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \left(\partial Q\left(\psi_s^{EG}(y_1 \dots y_{s-1}), y_s\right)\right)_k\right)}$$

(j-th component of $\psi_t^{EG}(y_1 \dots y_{t-1}) \in \mathbb{R}^N$)

⚠ ψ_t^{EG} is an extremely complicated function of its arguments $y_1 \dots y_{t-1}$ (cf. the various calls to ψ_s^{EG}, with $s \leq t-1$, in its definition).

We define $\tilde{\theta}_t = \psi_t^{EG}(y_{T+2-t}, \dots y_T)$ for $t \geq 2$

and $\tilde{\theta}_1 = \left(\tfrac{1}{N}, \dots \tfrac{1}{N}\right)$

We recommend: $\hat{\theta}_T = \frac{1}{T} \sum_{t=1}^{T} \tilde{\theta}_t$

⚠ Computing $\hat{\theta}_T$ is costly: the computation of each $\tilde{\theta}_t$ requires $t-1$ steps, namely the computation of the

$$\psi_s^{EG}\left(y_{T+2-t}, \dots y_{\underbrace{T-t}_{+s}}\right) \quad \text{for } 2 \leq s \leq t$$

no such intermediate quantity appears twice

$\Rightarrow$ about $\sum_{t=1}^{T} (t-1) = O(T^2/2)$ computation steps required.

* Proof of performance? (→ explains where the construction comes from !)

What does EG guarantee ?

$$\forall \theta \in \mathcal{X}, \quad \frac{1}{T}\sum_{t=1}^{T} Q\left(\psi_t^{EG}(y_1, \dots y_{t-1}),\ y_t\right) \leq \frac{1}{T}\sum_{t=1}^{T} Q(\theta, y_t) + \varepsilon_T$$

still $G\sqrt{(2\ln N)/T}$

We take expectations in both sides to conclude :

- Stationarity implies in particular that all the Y_t have the same distribution (but they are NOT independent in general); hence

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q(\theta, Y_t)\right] = \mathbb{E}\left[Q(\theta, Y_{T+1})\right]$$

- it also implies that $\left(\psi_t^{EG}(Y_1, \dots Y_{t-1}),\ Y_t\right)$ and $\left(\psi_t^{EG}(Y_{T+2-t}, \dots Y_T),\ Y_{T+1}\right) = \left(\tilde{\theta}_t,\ Y_{T+1}\right)$ have the same distribution, so that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q\left(\psi_t^{EG}(Y_1, \dots Y_{t-1}),\ Y_t\right)\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q(\tilde{\theta}_t, Y_{T+1})\right] \geq \mathbb{E}\left[Q(\hat{\theta}_T, Y_{T+1})\right]$$

↑ convexity of $\theta \mapsto Q(\theta, Y_{T+1})$

- thus we proved, for our choice of $\hat{\theta}_T$:

$$\mathbb{E}\left[Q(\hat{\theta}_T, Y_{T+1})\right] \leq \inf_{\theta \in \mathcal{X}} \mathbb{E}\left[Q(\theta, Y_{T+1})\right] + \varepsilon_T$$

↑ same ε_T as for the iid case.