Sequential setting #1 (meta-statistical)

* At each round $t = 1, 2, \dots T$:
  - experts output forecasts f_{jt}, $j \in \{1 \dots N\}$
  - statistician aggregates the forecasts as $\hat{y}_t = \sum_j p_{jt} f_{jt}$
    where $p_t = (p_{1t} \dots p_{Nt})$ is a convex weight vector
  - true observation y_t is revealed

* Assessment of performance: loss function $\ell(\hat{y}_t, y_t)$
  ↳ cumulative loss $\sum_{t=1}^{T} \ell(\hat{y}_t, y_t)$

No stochastic model → relative-performance criterion = perform almost as well as the best constant convex combination

ie, control $R_T = \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) - \inf_{q \in \chi} \sum_{t=1}^{T} \ell\left(\sum_{j=1}^{N} q_j f_{jt}, y_t\right)$

(χ: the simplex of all convex weight vectors; q: independent of t)

Ex: square loss $\ell(\hat{y}_t, y_t) = (\hat{y}_t - y_t)^2$
absolute loss $\ell(\hat{y}_t, y_t) = |\hat{y}_t - y_t|$
absolute percentage of error $\ell(\hat{y}_t, y_t) = |\hat{y}_t - y_t| / |y_t|$
↳ all these loss functions are convex in $\hat{y}_t$, a fact that we will need later on.

* R_T is called the regret

We have some bias/variance decomposition:

$$\sum_{t=1}^{T} \ell(\hat{y}_t, y_t) = \inf_{q} \sum_{t=1}^{T} \ell\left(\sum_j q_j f_{jt}, y_t\right) + R_T$$

- $\sum_{t=1}^{T} \ell(\hat{y}_t, y_t)$: cumulative loss
- $\inf_q \sum_{t=1}^{T} \ell\left(\sum_j q_j f_{jt}, y_t\right)$: approximation error (= how good the experts are)
- R_T: the regret corresponds to a sequential estimation error

We will focus on the regret in these lectures

Sequential setting #2 (sequential optimization)

* At each round $t = 1, 2, \dots T$:
  - Mathematician picks $p_t \in \mathcal{X}$
  - true loss function $\ell_t : \mathcal{X} \to \mathbb{R}$ is revealed

* Assessment of performance

$$R_T = \sum_{t=1}^{T} \ell_t(p_t) - \inf_{q \in \mathcal{X}} \sum_{t=1}^{T} \ell_t(q)$$

* Link : $\ell_t : p \mapsto \ell\left(\sum_j p_j f_{jt}, y_t\right)$

The only (small) difference is that the f_j are only available, in this version, AFTER p_t is selected.

* Special case of interest to start with : LINEAR loss functions

$$\ell_t(p) = \sum_{j=1}^{N} p_j \ell_{jt} \quad \text{where} \quad \ell_{jt} \overset{\text{not.}}{=} \ell_t(\delta_j)$$

↑ Dirac mass on j

ℓ_t is given by a vector $(\ell_{1t}, \dots \ell_{Nt})$, choosing ℓ_t is choosing such a vector.

Setting #2/
Linear losses → Exponentially weighted average predictor. [EWA]

* Algorithm: $p_1 = (1/N, \dots 1/N)$

$t \geq 2$, $p_{jt} = \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{js}\right) \Big/ \sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{ks}\right)$

formula also ok for $t=1$ since $\sum_{s=1}^{0} = 0$ by convention

where $\eta > 0$ is a learning rate.

Note: Not all weight on the best experts (it wouldn't work!) but most of the weight.

* Bound: For all strategies picking the linear loss functions $(\ell_{1t} \dots \ell_{Nt}) \in [m,M]^N$

$$R_T = \sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \ell_{jt} - \min_{q \in \Delta} \sum_{t=1}^{T} \sum_{j=1}^{N} q_j \ell_{jt}$$

$$= \sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \ell_{jt} - \min_{k=1\dots N} \sum_{t=1}^{T} \ell_{kt} \leq \frac{\ln N}{\eta} + \eta \frac{(M-m)^2}{8} T$$

PROOF: based on Hoeffding's lemma:

- for random variables $X \in [m,M]$,

$$\forall \eta \in \mathbb{R}, \quad \ln \mathbb{E}[e^{\eta X}] \leq \eta \mathbb{E}[X] + \frac{\eta^2}{8}(M-m)^2$$

- thus for all convex weight vectors $q \in \Delta$, for all $\ell_{jt} \in [m,M]$,

$$\forall \eta \in \mathbb{R}, \quad \ln \sum_j q_j e^{-\eta \ell_{jt}} \leq -\eta \sum_j q_j \ell_{jt} + \frac{\eta^2}{8}(M-m)^2$$

In particular, $\sum_j p_{jt} \ell_{jt} \leq -\frac{1}{\eta} \ln \sum_j p_{jt} e^{-\eta \ell_{jt}} + \frac{\eta}{8}(M-m)^2$

$$= \ln \frac{\sum_j \exp\left(-\eta \sum_{s=1}^{t} \ell_{js}\right)}{\sum_k \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{ks}\right)}$$

↑ by definition of the algorithm

Summing over t (telescoping sum in the right-hand side):

$$\sum_{t=1}^{T}\sum_{j=1}^{N} p_{jt}\,\ell_{jt} \le -\frac{1}{\eta}\ln\frac{\sum_{j=1}^{N}\exp\left(-\eta\sum_{t=1}^{T}\ell_{jt}\right)}{N} + \frac{\eta}{8}(M-m)^2 T$$

Conclude using $\ln\sum_j e^{-\eta\sum_{t=1}^{T}\ell_{jt}} \ge \ln\max_j e^{-\eta\cdots} = \max_j \ln e^{-\eta\cdots} = \max_j\{-\eta\cdots\}$

$$= -\eta\min_j\sum_{t=1}^{T}\ell_{jt}$$

and (linearity): $\min_{j=1..N}\sum_{t=1}^{T}\ell_{jt} = \min_{q\in\mathcal{X}}\sum_{t=1}^{T}\sum_{k} q_k\,\ell_{kt}$. ⌟

Reminder:

⌜ PROOF of HOEFFDING'S LEMMA:

$\Psi(\eta) = \ln\mathbb{E}[e^{\eta X}]$ defined for all $\eta\in\mathbb{R}$

$\Psi'(\eta) = \dfrac{\mathbb{E}[Xe^{\eta X}]}{\mathbb{E}[e^{\eta X}]}$ (differentiability: cf. X bounded thus $\eta\mapsto Xe^{\eta X}$ locally dominated by an integrable r.v. independent of η)

(Ψ twice differentiable: similar reasons)

$$\Psi''(\eta) = \frac{\mathbb{E}[X^2e^{\eta X}]\mathbb{E}[e^{\eta X}] - \left(\mathbb{E}[Xe^{\eta X}]\right)^2}{\left(\mathbb{E}[e^{\eta X}]\right)^2} = \mathrm{Var}_{\mathbb{Q}}(X)$$

under the probability $\mathbb{Q}$ defined by

$$\frac{d\mathbb{Q}}{d\mathbb{P}}(\omega) = \frac{e^{\eta X(\omega)}}{\mathbb{E}[e^{\eta X}]}$$

$X\in[m,M]$: $\mathrm{Var}_{\mathbb{Q}}(X) = \inf_{\mu\in\mathbb{R}}\mathbb{E}_{\mathbb{Q}}\left[(X-\mu)^2\right] \le \mathbb{E}_{\mathbb{Q}}\left[\left(X-\frac{M+m}{2}\right)^2\right] \le \frac{(M-m)^2}{4}$

Taylor: $\exists x$ s.t. $\Psi(\eta) = \overbrace{\Psi(0)}^{=0} + \eta\,\Psi'(0) + \frac{\eta^2}{2}\overbrace{\Psi''(x)}$

(cf. Ψ is actually C^2)

ie, $\ln\mathbb{E}[e^{\eta X}] \le \eta\,\mathbb{E}[X] + \frac{\eta^2}{8}(M-m)^2$ ⌟

* How good is the obtained bound?

$$\min_{\eta>0}\ \frac{\ln N}{\eta} + \frac{\eta}{8}(M-m)^2T = (M-m)\sqrt{\frac{T}{2}\ln N}$$

nice because $o(T)$ and weak dependence in N

for $\eta^* = \dfrac{1}{(M-m)}\sqrt{\dfrac{8\ln N}{T}}$ ← not nice when T, m, M unknown

Why bother? Why not put all mass on the best j in EWA?
Why only compare the performance to the best global expert?
How good is EWA?

(exercise #1)

1) Consider the strategy

$$p_1 = \left(\tfrac{1}{N} \dots \tfrac{1}{N}\right)$$

$$p_t \longleftrightarrow \begin{cases} \text{equal mass on } k \text{ s.t. } \sum_{s\le t-1} \ell_{ks} = \min_j \sum_{s\le t-1} \ell_{js} \\ \text{zero mass on } k \text{ s.t. } \sum_{s\le t-1} \ell_{ks} > \min_j \sum_{s\le t-1} \ell_{js} \end{cases}$$

called: « follows the leader(s) » (FTL)

Show that it fails: these exist sequences $(\ell_{1t} \dots \ell_{Nt}) \in [0,1]^N$ such that $\exists \delta > 0 \mid \forall T$, $R_T = \sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \ell_{jt} - \min_{k=1\dots N} \sum_{t=1}^{T} \ell_{kt} \ge \delta T$.

EWA appears as a smoothed version of FTL!

2) Can we be more ambitious? People often hope to get close to $\sum_{t=1}^{T} \min_{k=1\dots N} \ell_{kt}$ in statistics when they resort to aggregation.

Show that no strategy can be such that for all sequences $(\ell_{1t} \dots \ell_{Nt}) \in [0,1]^N$, $R_T = \sum_{t,j} p_{jt} \ell_{jt} - \sum_{t} \min_k \ell_{kt} = o(T)$.

3) Show that for EWA, the regret is never negative:

$\forall (\ell_{1t} \dots \ell_{Nt}) \in [m, M]^N$, $R_T = \sum_{jt} p_{jt} \ell_{jt} - \min_{k=1\dots N} \sum \ell_{kt} \ge 0$

REMINDER: When does a limit in distribution entail a limit for the expectations?

(1) In general, $Y_T \xrightarrow{\mathcal{L}} Y$ does not imply $\mathbb{E}[f(Y_T)] \to \mathbb{E}[f(Y)]$ if f is just assumed continuous.

Ex: $f = \text{identity}$, $Y_T = T^2 \mathbb{1}_{\{U_T \in [0,1/T]\}}$ where $U_1 \dots U_T \dots$ iid $\sim \mathcal{U}_{[0,1]}$

We have $\begin{cases} Y_T \xrightarrow{\mathbb{P}} 0 \\ \mathbb{E}[Y_T] = T \longrightarrow +\infty \end{cases}$

(2) Definition of $\xrightarrow{\mathcal{L}}$ is for continuous and bounded functions:

Let $(Y_T)_{T\geq 1}$ be a sequence of $\mathbb{R}^N$-valued random variables

$Y_T \xrightarrow{\mathcal{L}} Y$ if by definition, for all continuous and bounded functions $f: \mathbb{R}^N \to \mathbb{R}$, we have

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

(3) Uniform asymptotic integrability [uai]: is a sufficient condition for (1) to hold

Def: $(Z_T)_{T\geq 1}$ is uai if

$$\lim_{L\to+\infty} \limsup_{T\to+\infty} \mathbb{E}\left[\|Z_T\| \mathbb{1}_{\{\|Z_T\|>L\}}\right] = 0$$

Lemma: If $(Y_T)_{T\geq 1}$ is a sequence of random vectors in $\mathbb{R}^N$ and $f: \mathbb{R}^N \to \mathbb{R}$ is a continuous function such that

- $Y_T \xrightarrow{\mathcal{L}} Y$
- $(f(Y_T))_{T\geq 1}$ is uai,

then: a) $f(Y) \in \mathbb{L}^1$ and $f(Y_T) \in \mathbb{L}^1$ for all T sufficiently large

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

Proof: For a): 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, $\limsup_{T\to+\infty} \mathbb{E}[|f(Y_T)|] < +\infty$,

thus, $\left(\mathbb{E}[|f(Y_T)|]\right)_{T \geq T_0}$ is bounded for some T_0

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

$$f(Y_T) \longrightarrow f(Y) \quad \text{as}$$

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

For b): We denote by φ_L : $x \in \mathbb{R} \mapsto \max\{-L, \min\{x, L\}\}$ the clipping operator

$$\begin{aligned} |\mathbb{E}[f(Y)] - \mathbb{E}[f(Y_T)]| \leq\ & |\mathbb{E}[f(Y)] - \mathbb{E}[\varphi_L(f(Y))]| \\ & + |\mathbb{E}[\varphi_L(f(Y))] - \mathbb{E}[\varphi_L(f(Y_T))]| \\ & + |\mathbb{E}[\varphi_L(f(Y_T))] - \mathbb{E}[f(Y_T)]| \end{aligned}$$

$\forall L$: tends to 0 as $Y_T \xrightarrow{\mathcal{L}} Y$ and $\varphi_L \circ f$ is continuous and bounded

Thus,

$$\begin{aligned} \forall L,\ \limsup_{T \to +\infty} & |\mathbb{E}[f(Y)] - \mathbb{E}[f(Y_T)]| \\ & \leq |\mathbb{E}[f(Y)] - \mathbb{E}[\varphi_L(f(Y))]| \\ & + \limsup_{T \to +\infty} |\mathbb{E}[\varphi_L(f(Y_T))] - \mathbb{E}[f(Y_T)]| \end{aligned}$$

Taking $\limsup_{L \to +\infty}$ in both sides (well, the left-hand side is independent of L) and using that $\limsup_{T \to +\infty} a_T + b_T \leq \limsup_{T \to +\infty} a_T + \limsup_{T \to +\infty} b_T$ for non-negative sequences (a_T) and (b_T):

$$\begin{aligned} \limsup_{T \to +\infty} & |\mathbb{E}[f(Y)] - \mathbb{E}[f(Y_T)]| \\ & \leq \underbrace{\limsup_{L \to +\infty} |\mathbb{E}[f(Y)] - \mathbb{E}[\varphi_L(f(Y))]|}_{= 0 \text{ by dominated convergence}} + \underbrace{\limsup_{L \to +\infty} \limsup_{T \to +\infty} |\mathbb{E}[f(Y_T)] - \mathbb{E}[\varphi_L(f(Y_T))]|}_{\text{to be handled by the weak property}} \end{aligned}$$

By definition of φ_L,

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

In particular, $\left|\mathbb{E}[f(Y_T)] - \mathbb{E}[\varphi_L(f(Y_T))]\right|$

$$\leqslant \mathbb{E}\left[\underbrace{|f(Y_T)+L|}_{\leqslant |f(Y_T)| \text{ on this event}}\,\mathbb{1}_{\{f(Y_T)<-L\}}\right] + \mathbb{E}\left[\underbrace{|f(Y_T)-L|}_{\leqslant |f(Y_T)| \text{ on this event}}\,\mathbb{1}_{\{f(Y_T)>L\}}\right]$$

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

whose $\limsup_{L\to+\infty} \limsup_{T\to+\infty}$ ... equals 0 by the very definition of uai for $(f(Y_T))_{T\geqslant 1}$

This concludes the proof.

(4) A useful sufficient condition for uai:

Lemma: If $(Z_T)_{T\geqslant 1}$ is bounded in L^p for some $p>1$, ie $\sup_{T\geqslant 1} \mathbb{E}[\|Z_T\|^p] \overset{\text{not.}}{=} B < +\infty$.

Then $(Z_T)_{T\geqslant 1}$ is uai.

Proof: $\varphi(x) = x^p$: as $p>1$, $\varphi(x) \gg x$

Thus, $\forall M>0,\ \exists L \mid \forall x \geqslant L,\ \varphi(x) \geqslant Mx$

Taking $M = B\gamma$: $\exists L_\gamma \mid \forall T\geqslant 1,\ B \geqslant \mathbb{E}[\|Z_T\|^p] \geqslant \mathbb{E}\left[\|Z_T\|^p\,\mathbb{1}_{\{\|Z_T\|>L\}}\right]$

$$\geqslant B\gamma\,\mathbb{E}\left[\|Z_T\|\,\mathbb{1}_{\{\|Z_T\|>L_\gamma\}}\right]$$

Thus, for this L: $\sup_{T\geqslant 1} \mathbb{E}\left[\|Z_T\|\,\mathbb{1}_{\{\|Z_T\|>L_\gamma\}}\right] \leqslant \frac{1}{\gamma}$

and the uai property follows by letting $\gamma \to +\infty$.
(even something stronger than that follows)

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 $\ell_{jt} \in [0,1]$ for all t and j (ie, $m=0$ and $M=1$)

Theorem:

$$\liminf_{N\to+\infty} \liminf_{T\to+\infty} \inf_{\text{algorithms}} \sup_{\ell_{jt}\in[0,1]} \frac{\sum_{t,j} p_{jt}\ell_{jt} - \min_k \sum_t \ell_{kt}}{\sqrt{\frac{T}{2}\ln N}} \geq 1$$

↑ stronger to have an infimum here than to probe $\liminf \liminf \sup \cdots \geq 1$ $\forall$ algo,

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

Proof: We lower bound the $\sup_{\ell_{jt}\in[0,1]}$ by an $\mathbb{E}[\ ]$ with losses ℓ_{jt} drawn iid $\sim \mathrm{Ber}(1/2)$

Using that for any algorithm:

$$\forall t \geq 1, \quad \mathbb{E}\Big[\sum_j p_{jt}\ell_{jt}\Big] = \sum_j \mathbb{E}[p_{jt}]\,\underbrace{\mathbb{E}[\ell_{jt}]}_{=1/2} = \frac{1}{2}\mathbb{E}\Big[\underbrace{\sum_j p_{jt}}_{=1 \text{ a.s.}}\Big] = \frac{1}{2}$$

↑ p_{jt} is measurable wrt past losses ℓ_{ks} $s \leq t-1$ and is thus independent of the ℓ_{jt}

↳ Not so surprising: means that we cannot predict efficiently fair coin tosses

* But * by a crowd effect (due to the central limit theorem) we will show that

$$\mathbb{E}\Big[\min_k \sum_{t=1}^T \ell_{kt}\Big] = T/2 - \text{something of order } \sqrt{T\ln N}$$

even though we "individually" have $\forall k,\ \mathbb{E}\Big[\sum_{t=1}^T \ell_{kt}\Big] = T/2$

More formally:

$$\inf_{\text{algo}} \sup_{\ell_{jt}\in[0,1]} \frac{\sum_{t,j} p_{jt}\ell_{jt} - \min_k \sum_t \ell_{kt}}{\sqrt{\frac{T}{2}\ln N}} \geq \inf_{\text{algo}} \frac{\mathbb{E}\Big[\sum_{t,j} p_{jt}\ell_{jt}\Big] - \mathbb{E}\Big[\min_k \sum_{t=1}^T \ell_{kt}\Big]}{\sqrt{(T/2)\ln N}}$$

$$= \frac{1}{\sqrt{2\ln N}}\,\mathbb{E}\left[\max_{k=1\dots N} \frac{\sum_{t=1}^T (1/2 - \ell_{kt})}{\frac{1}{2}\sqrt{T}}\right]$$

We denote $Z_{kT} = \dfrac{\sum_{t=1}^{T} \left(\frac{1}{2} - L_{kt}\right)}{\frac{1}{2}\sqrt{T}}$ ← sum of iid random var ← $\sigma\sqrt{T}$

and note that Z_{kT} is a central-limit-Theorem statistic: $\frac{1}{2} - L_{kt}$ are iid, centered with $\mathrm{Var}\left(\frac{1}{2} - L_{kt}\right) = 1/4$

For each k, $Z_{kT} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$

By independence, $Z_T = \begin{pmatrix} Z_{1T} \\ \vdots \\ Z_{NT} \end{pmatrix} \xrightarrow{\mathcal{L}} Z \sim \mathcal{N}(0, I_N)$

Take $f_N(x_1, \dots x_N) = \max_{k \le N} x_k$; we have shown so far:

$$\inf_{\text{algo}} \sup_{\ell_{jt} \in [0,1]} \frac{\sum_{t,j} p_j \ell_{jt} - \min_k \sum_t \ell_{kt}}{\sqrt{\frac{T}{2} \ln N}} \geq \frac{1}{\sqrt{2 \ln N}} \mathbb{E}\left[f_N(Z_T)\right]$$

The $f_N(Z_T)$ are bounded by N in L^2-norm:

$$\mathbb{E}\left[f_N(Z_T)^2\right] = \mathbb{E}\left[\max_{k \le N} Z_{kT}^2\right] \le \sum_{k=1}^{N} \underbrace{\mathbb{E}\left[Z_{kT}^2\right]}_{\text{each} = 1, \text{ cf. } Z_{kT} \text{ central-limit-Theorem statistic}} = N$$

Thus they form a uai sequence and we have

$$\lim_{T \to +\infty} \frac{1}{\sqrt{2 \ln N}} \mathbb{E}\left[f_N(Z_T)\right] = \frac{1}{\sqrt{2 \ln N}} \mathbb{E}\left[\max_{k \le N} Z_k\right]$$

So far:

$$\liminf_{T \to +\infty} \inf_{\text{algo}} \sup_{\ell_{jt} \in [0,1]} \frac{\sum_{t,j} p_j \ell_{jt} - \min_k \sum_t \ell_{kt}}{\sqrt{(T/2) \ln N}} \geq \frac{1}{\sqrt{2 \ln N}} \mathbb{E}\left[\max_{k \le N} Z_k\right]$$

The proof is concluded by: ↑ details on the next pages

$$\liminf_{N \to +\infty} \dots \geq 1$$

(this is even a true limit)

Asymptotics for the expectation of the maximum of standard Gaussian random variables

We prove that for $G_1, G_2, \dots$ iid $\sim \mathcal{N}(0,1)$,

$$\mathbb{E}\left[\max_{k \le N} G_k\right] \sim \sqrt{2 \ln N}$$

(1) Easy part (useless for the main result proved above):

$$\mathbb{E}\left[\max_{k \le N} G_k\right] \le \sqrt{2 \ln N}$$

as a special case of the following lemma.

Lemma: If $V_1, V_2, \dots, V_N$ are such that $\forall j \ \forall \lambda > 0 \quad \mathbb{E}\left[e^{\lambda V_j}\right] \le e^{\lambda^2 \sigma^2 / 2}$

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

Notes:

- Valid even without independence (independence will be needed for the lower bound)
- If $G_k \sim \mathcal{N}(0, \sigma^2)$ then $\mathbb{E}\left[e^{\lambda G_k}\right] = e^{\lambda^2 \sigma^2 / 2}$
- If $G_k \in [m, M]$ with $\mathbb{E}[G_k] = 0$ then by Hoeffding's lemma $\mathbb{E}\left[e^{\lambda G_k}\right] \le e^{\lambda^2 (M-m)^2 / 8}$
- We call sub-Gaussian the random variables V_k satisfying the assumptions of the lemma.

Proof: Proof technique called "Pisier's argument":

$$\mathbb{E}\left[\max_{k \le N} V_k\right] \le \frac{1}{\lambda} \ln \mathbb{E}\left[\exp\left(\lambda \max_{k \le N} V_k\right)\right] = \frac{1}{\lambda} \ln \mathbb{E}\left[\max_{k \le N} e^{\lambda V_k}\right]$$

Jensen's inequality

$$\le \frac{1}{\lambda} \ln\left(\sum_{k \le N} \mathbb{E}\left[e^{\lambda V_k}\right]\right)$$

max of non-neg quantities $\le$ their sum

$$\le \frac{1}{\lambda} \ln\left(N e^{\lambda^2 \sigma^2 / 2}\right) = \frac{\ln N}{\lambda} + \frac{\lambda \sigma^2}{2}$$

by the sub-Gaussian assumption

$$= \sqrt{2\sigma^2 \ln N}$$

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

(2) "Difficult" part (and the one that we need!):

$$\liminf_{N\to+\infty} \frac{\mathbb{E}[\max_{k\le N} G_k]}{\sqrt{2\ln N}} \geq 1$$

Credits: Pascal Massart + Nathieu Even, M2 student following this course in 2019-20

$$M_N \overset{\text{def.}}{=} \left(\max_{k\le N} G_k \Big/ \sqrt{2\ln N}\right)^+ \quad \leftarrow \text{non-negative part}$$

Given that $0 \geq \mathbb{E}\left[\left(\max_{k\le N} G_k\right) \mathbb{1}_{\{\max_{k\le N} G_k \le 0\}}\right]$

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

$$\geq_{\text{(Cauchy-Schwarz)}} -\sqrt{\mathbb{E}[G_1^2]\,\mathbb{P}\{G_1\le 0\}} = -1/\sqrt{2}$$

we have

$$\liminf_{N\to+\infty} \frac{\mathbb{E}[\max_{k\le N} G_k]}{\sqrt{2\ln N}} = \liminf_{N\to+\infty} \mathbb{E}[M_N]$$

By Fatou's lemma, it thus suffices to show that $\liminf_{N\to+\infty} M_N \geq 1$ a.s.

For $\varepsilon \in (0,1)$:

$$\mathbb{P}\left(M_N \le \sqrt{1-\varepsilon}\right) = \mathbb{P}\left(\forall k\le N,\ G_k \le \sqrt{2(1-\varepsilon)\ln N}\right)$$

(the positive part in the def. of M_N)

$$= \left(\Phi\left(\sqrt{2(1-\varepsilon)\ln N}\right)\right)^N \le \exp\left(-N\left(1-\Phi(\dots)\right)\right)$$

(thus $\le e^a$)

where Φ is the cdf of the $\mathcal{N}(0,1)$ distribution.

As $1-\Phi(x) \underset{x\to+\infty}{\sim} e^{-x^2/2}/x\sqrt{2\pi}$ (proof by two integrations by parts, see below)

we have

$$N\left(1-\Phi\left(\sqrt{2(1-\varepsilon)\ln N}\right)\right) \sim N\exp\left(-(1-\varepsilon)\ln N\right) \Big/ 2\sqrt{\pi(1-\varepsilon)\ln N}$$

$$= N^{\varepsilon}/2\sqrt{\pi(1-\varepsilon)\ln N}$$

Thus $\sum_N \exp\left(-N\left(1-\Phi(\ )\right)\right)$ is convergent

therefore, $\sum_N \mathbb{P}\left(M_N \le \sqrt{1-\varepsilon}\right)$ is also convergent

By the Borel-Cantelli lemma: $\{M_N \leq \sqrt{1-\varepsilon}\}$ a.s. takes place only finitely many times,

ie $\liminf_{N\to\infty} M_N \geq \sqrt{1-\varepsilon}$ a.s.

Since this is true for all $\varepsilon>0$, we get the desired $\liminf_{N\to\infty} M_N \geq 1$ a.s. (with $\varepsilon_n = 1/n \to 0$)

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

First IbP (integration by parts):

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

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

Second IbP:

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

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

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

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

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 iid $\sim \mathcal{B}(1/2)$ $\forall k,t$

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

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

(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:

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

(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-processing 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 $a \geqslant 2e/2e+1$

Show that $KL(P_{\mathbb{E}_1}, P_{\mathbb{E}_2}) \leq KL(P_{\mathbb{Q}_1}, P_{\mathbb{Q}_2})$ for any random variables $\mathbb{E}_1, \mathbb{E}_2$ (defined on the same probability space) and any probability distributions $\mathbb{Q}_1, \mathbb{Q}_2$ on the same space, with equality if $\mathbb{E}_1 = \mathbb{E}_2$ almost surely.

Hint: →

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

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

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 $N \geqslant 2$ and $T \geqslant 17 \ln N$,

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

Application of the EWA forecaster / Sion's lemma.

Statement: Let $\mathcal{X}, \mathcal{Y}$ two convex sets, $f: \mathcal{X} \times \mathcal{Y} \to [0, M]$ a function s.t. $\forall x \in \mathcal{X}$, $f(x, \cdot)$ is concave, $\forall y \in \mathcal{Y}$, $f(\cdot, y)$ is convex

then (under additional regularity assumptions):

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x,y) = \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x,y).$$

Proof:

1) $\geq$ always holds: $\forall x, \forall y$, $f(x,y) \geq \inf_{x' \in \mathcal{X}} f(x', y)$

taking $\sup_{y \in \mathcal{Y}}$ in both sides: $\forall x$, $\sup_{y \in \mathcal{Y}} f(x,y) \geq \sup_{y \in \mathcal{Y}} \inf_{x' \in \mathcal{X}} f(x', y)$

Get $\inf \sup f \geq \sup \inf f$ by taking the $\inf_{x \in \mathcal{X}}$ (the right-hand side is a constant independent of x).

2) A (fictitious) statistician and a (fictitious) opponent play as follows:

First, the statistician sets $N \geq 2$ and $x^{(1)}, \dots, x^{(N)}$ in $\mathcal{X}$, as well as $T \geq 1$

Then, at each round, they simultaneously pick

$$x_t = \sum_{j=1}^{N} p_{jt} x^{(j)} \in \mathcal{X} \quad \text{and} \quad y_t \in \mathcal{Y}$$

with $\eta = \frac{1}{M}\sqrt{\frac{8 \ln N}{T}}$

How? $p_{jt} = \exp\left(-\eta \sum_{s=1}^{t-1} f(x^{(j)}, y_s)\right) \Big/ \sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} f(x^{(k)}, y_s)\right)$

and (since p_{jt} only depends on the past, the opponent can compute it and pick:)

y_t s.t. (by definition of sup) $f(x_t, y_t) \geq \sup_{y \in \mathcal{Y}} f(x_t, y) - 1/T$

By definition of the exponentially weighted average strategy with a well-chosen learning rate $\eta > 0$

$$\sum_{t=1}^{T} \sum_{j=1}^{N} p_{j,t} \underbrace{f(x^{(j)}, y_t)}_{\text{corresponds to } g_t \in [0,M]} - \min_{k=1,\dots N} \sum_{t=1}^{T} f(x^{(k)}, y_t) \leq \underbrace{\frac{\ln N}{\eta} + \eta \frac{M^2}{8} T}_{= M\sqrt{\frac{T}{2}\ln N}}$$

given your choice for η

From the convexity of $f(\cdot, y_t)$, we finally get:

$$\sum_{t=1}^{T} f\Big(\underbrace{\sum_j p_{j,t} x^{(j)}}_{= x_t \text{ by def.}}, y_t\Big) - \min_{k=1\dots N} \sum_{t=1}^{T} f(x^{(k)}, y_t) \leq M\sqrt{\frac{T}{2}\ln N} \quad (*)$$

where $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$

3) Now, $\inf_x \sup_y f(x,y) \leq \sup_y f(\bar{x}, y)$

$\leq \sup_y \frac{1}{T}\sum_{t=1}^{T} f(x_t, y)$ ($f(\cdot,y)$ convex $\forall y$)

$\leq \frac{1}{T}\sum_{t=1}^{T} \sup_y f(x_t, y)$ ($\sup \Sigma \leq \Sigma \sup$)

$\leq \frac{1}{T}\sum_{t=1}^{T} f(x_t, y_t) + 1/T$ (def of y_t)

while by (*):

$$\frac{1}{T}\sum_{t=1}^{T} f(x_t, y_t) \leq \underbrace{M\sqrt{\frac{\ln N}{2T}}}_{= O(1/\sqrt{T})} + \min_{k=1\dots N} \frac{1}{T}\sum_{t=1}^{T} f(x^{(k)}, y_t)$$ (by (*))

$$\leq O(1/\sqrt{T}) + \min_k f(x^{(k)}, \bar{y})$$ (by concavity of $f(x^{(k)}, \cdot)$, where $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$)

$$\leq \sup_{y \in \mathcal{Y}} \min_k f(x^{(k)}, y)$$

4) In sections (2)-(3) T, N and $x^{(1)} \dots x^{(N)}$ were fixed, but we can play with them! We proved

$$\inf_x \sup_y f(x,y) \leq \frac{1}{T} + O(1/\sqrt{T}) + \sup_y \min_k f(x^{(k)}, y)$$

Letting $T \to +\infty$:

$$\inf_x \sup_y f(x,y) \leq \sup_y \min_k f(x^{(k)},y)$$

This holds for all N and all $x^{(1)} \dots x^{(N)}$ in $\mathcal{X}$:

$$\inf_x \sup_y f(x,y) \leq \inf_{N\geq 1} \inf_{\{x^{(1)}\dots x^{(N)}\}\subset\mathcal{X}} \sup_{y\in\mathcal{Y}} \min_{k=1\dots N} f(x^{(k)},y)$$

$$\geq \sup_y \inf_x f(x,y) \quad \text{but maybe} \leq \sup_y \inf_x f(x,y)$$

is clearly ?

so that we have an equality?

We will now state and use regularity / topological assumptions.

Assume
- $\mathcal{X}, \mathcal{Y}$ are metric spaces, with distances $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$
- $f : \mathcal{X}\times\mathcal{Y} \to [a,M]$ uniformly continuous :

(In particular true when $\mathcal{X}$ is included in a compact set)

$$\forall \varepsilon>0,\ \exists \delta>0 \mid d_{\mathcal{X}}(x,x') + d_{\mathcal{Y}}(y,y') \leq \delta \Rightarrow |f(x,y) - f(x',y')| \leq \varepsilon$$

- Finite covering property for $\mathcal{X}$:

$$\forall \delta>0,\ \exists N \text{ and } x^{(1)}\dots x^{(N)} \text{ s.t. } \mathcal{X} \subset \bigcup_{j=1}^{N} B(x^{(j)},\delta)$$

Given $\varepsilon>0$ and the associated $\delta>0$:

$$\forall x,y,\quad f(x,y) \geq \min_{j=1\dots N} f(x^{(j)},y) - \varepsilon$$ by uniform continuity

Taking $\inf_x$ then $\sup_y$:

$$\sup_y \inf_x f(x,y) \geq \sup_y \min_j f(x^{(j)},y) - \varepsilon$$

$$\geq \inf_N \inf_{\{x^{(j)}\}} \sup_y \min_j f(x^{(j)},y) - \varepsilon$$

and we let $\varepsilon \downarrow 0$

EXERCISE #2
Bonus points reward!

(Perhaps you can find even better — weaker — assumptions? If so, let me know! ↑ while still having a smooth and easy-to-read proof...)