Part 3: Fano-type inequalities.

Simplest framework: measurable space $(\Omega, \mathcal{F})$

finitely many probability distributions $\mathbb{P}_i, \mathbb{Q}_i$ $i \in \{1, \dots N\}$

events $A_i \in \mathcal{F}$

Aim: bound $\frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_i(A_i)$ by $\frac{1}{N} \sum_{i=1}^{N} \mathbb{Q}_i(A_i) + \dots$

Most general framework: underlying measurable space $(\Omega, \mathcal{F})$

parameter space $(\Theta, \mathcal{G})$ with prior distribution ν

families of probability distributions $\mathbb{P}_\theta, \mathbb{Q}_\theta$ $\theta \in \Theta$

family of $[q]$-valued random variables Z_θ $\theta \in \Theta$

(+ measurability assumptions: $\theta \mapsto \mathbb{E}_{\mathbb{P}_\theta}[Z_\theta]$ and $\theta \mapsto \mathbb{E}_{\mathbb{Q}_\theta}[Z_\theta]$ are $\mathcal{G}$-measurable)

Aim: bound $\int_\Theta \mathbb{E}_{\mathbb{P}_\theta}[Z_\theta] \, d\nu(\theta)$ by $\int_\Theta \mathbb{E}_{\mathbb{Q}_\theta}[Z_\theta] \, d\nu(\theta) + \dots$

Two steps:
1. Reduction to Bernoulli distributions
2. Lower bound the kl function

Reminder of a short-hand notation: $kl(p,q) = KL(Ber(p), Ber(q))$

Step 1:

$$kl\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_i(A_i), \frac{1}{N}\sum_{i=1}^{N} \mathbb{Q}_i(A_i)\right) \leq \frac{1}{N}\sum_{i=1}^{N} kl(\mathbb{P}_i(A_i), \mathbb{Q}_i(A_i))$$

$$\leq \frac{1}{N}\sum_{i=1}^{N} KL(\mathbb{P}_i, \mathbb{Q}_i)$$

(convexity of KL; data-processing inequality)

$$kl\left(\int_\Theta \mathbb{E}_{\mathbb{P}_\theta}[Z_\theta] \, d\nu(\theta), \int_\Theta \mathbb{E}_{\mathbb{Q}_\theta}[Z_\theta] \, d\nu(\theta)\right) \leq \int_\Theta kl(\mathbb{E}_{\mathbb{P}_\theta}[Z_\theta], \mathbb{E}_{\mathbb{Q}_\theta}[Z_\theta]) \, d\nu(\theta)$$

$$\leq \int_\Theta KL(\mathbb{P}_\theta, \mathbb{Q}_\theta) \, d\nu(\theta)$$

(under the needed measurability assumptions)

In both cases: $kl(\bar{p}, \bar{q}) \leq \bar{K}$

Step 2:

$$\begin{cases} kl(p,q) \geq p \ln \frac{1}{q} - \ln 2 \\ kl(p,q) \geq \left(\ln \frac{1}{q}\right)(p-q)^2 \end{cases} \quad \text{for all } (p,q) \in [0,1]^2$$

as we saw last time.

Thus:

$$\begin{cases} p \leq \dfrac{\ln 2 + kl(p,q)}{\ln(1/q)} \\ p \leq q + \sqrt{\dfrac{kl(p,q)}{\ln(1/q)}} \end{cases} \quad \text{for all } p \in [0,1] \text{ and } q \in (0,1)$$

Conclusions:

E.g.: (1) $$\frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_i(A_i) \leq \frac{\ln 2 + \frac{1}{N}\sum_{i=1}^{N} KL(\mathbb{P}_i, \mathbb{Q}_i)}{-\ln\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{Q}_i(A_i)\right)}$$

(2) $$\frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_i(A_i) \leq \frac{1}{N}\sum_{i=1}^{N} \mathbb{Q}_i(A_i) + \sqrt{\frac{\sum_{i=1}^{N} KL(\mathbb{P}_i, \mathbb{Q}_i)}{-\ln\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{Q}_i(A_i)\right)}}$$

(3) $$\int_\Theta \mathbb{E}_{\mathbb{P}_\theta}[Z_\theta]\, d\nu(\theta) \leq \int_\Theta \mathbb{E}_{\mathbb{Q}_\theta}[Z_\theta]\, d\nu(\theta) + \sqrt{\frac{\int_\Theta KL(\mathbb{P}_\theta, \mathbb{Q}_\theta)\, d\nu(\theta)}{-\ln \int_\Theta \mathbb{E}_{\mathbb{Q}_\theta}[Z_\theta]\, d\nu(\theta)}}$$

Comments:

If the A_i form a partition of Ω and $\mathbb{Q}_i = \mathbb{Q}$ $\forall i$:

(1) says: $$\frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_i(A_i) \leq \frac{\ln 2 + \frac{1}{N}\sum_{i=1}^{N} KL(\mathbb{P}_i, \mathbb{Q})}{\ln N}$$

→ non-trivial only if $N \geq 3$

Birgé's (2005) version: $$\min_{i=1..N} \mathbb{P}_i(A_i) \leq \max\left\{0.84, \frac{\frac{1}{N-1}\sum_{i=2}^{N} KL(\mathbb{P}_i, \mathbb{P}_1)}{\ln N}\right\}$$

Alternative idea: resort to (2)!

which indicates $$\frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_i(A_i) \leq \frac{1}{N} + \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N} KL(\mathbb{P}_i, \mathbb{Q})}{\ln N}}$$

→ Which Q should be chosen in the case $Q_i = Q \ \forall i$?

Left-hand side independent of Q

Right-hand side: minimize $\frac{1}{N} \sum_{i=1}^{N} KL(P_i, Q)$

See Yang and Barron (1999):

We may restrict our attention to Q s.t. $P_i \ll Q \ \forall i$

then $\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i$ is s.t. $\forall j, \ P_j \ll \bar{P} \ll Q$

in particular, Q-as: $\frac{dP_j}{dQ} = \frac{dP_j}{d\bar{P}} \frac{d\bar{P}}{dQ}$

so that

$$\frac{1}{N} \sum_{j=1}^{N} KL(P_j, Q) = \frac{1}{N} \sum_{j=1}^{N} \int \left(\ln \frac{dP_j}{d\bar{P}} + \ln \frac{d\bar{P}}{dQ} \right) dP_j$$

$$= \frac{1}{N} \sum_{j=1}^{N} KL(P_j, \bar{P}) + KL(\bar{P}, Q)$$

The infimum over Q is achieved at $Q = \bar{P}$.

Application #1: Robust sequential learning with sparse losses.

Setting: For each $t = 1, 2, \dots$

- The environment picks $(\ell_{1t}, \dots \ell_{Nt}) \in [0,1]^N$ while simultaneously, the decision maker picks $I_t \in \{1, \dots N\}$, possibly at random thanks to $U_t \sim \mathcal{U}_{[0,1]}$ where $U_1, U_2, \dots$ independent
- Both players observe $(\ell_{1t}, \dots \ell_{Nt})$ and I_t

Aim of the decision-maker: minimize, ie, control from above, the regret:

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

Remark: « Robust sequential learning » because the ℓ_{jt} are not realizations of some well-behaved stochastic process, they can be picked by an opponent to annoy the decision maker

See the monograph by Cesa-Bianchi and Lugosi (2006).

Variant / generalization by Kwon and Perchet (2015):

Only s components of $(\ell_{1t}, \dots \ell_{Nt})$ are non-zero, where $s \in \{1, \dots N\}$ is fixed and known to the decision maker.

Denote by $H_{t-1} = \left((\ell_{js})_{j \in \{1\dots N\}}, U_s \right)_{s \in \{1, \dots t-1\}}$ the information available to the decision-maker when picking the law of I_t.

Theorem: There exists a strategy for the decision maker [namely, the exponentially weighted average algorithm with a well-chosen time-adaptive learning rate $(\eta_t)_t$] such that: for all $T \geq 1$, for all strategies of the environment leading to losses $\ell_{jt} \in [0,1]$ such that $(\ell_{1t} \dots \ell_{Nt})$ has at most s non-null components,

$$\bar{R}_T = \sum_{t=1}^{T} \mathbb{E}\left[\ell_{I_t,t} \mid \sigma(H_{t-1})\right] - \min_{k=1\dots N} \sum_{t=1}^{T} \ell_{kt} \leq C\left(\sqrt{sT \frac{\ln N}{N}} + \ln N\right)$$

where C is some numerical constant.

(In Kwon and Perchet, 2015 they use a fixed η tuned as a function of N, s and T and get the result for $C = \sqrt{2}$.)

Consequences:

* The Hoeffding-Azuma inequality ensures that:

$\forall T$, with probability at least $1-\delta$,

$$R_T = \sum_{t=1}^{T} \ell_{I_t,t} - \min_{k=1..N} \sum_{t=1}^{T} \ell_{kt} \leq C\left(\sqrt{sT\frac{\ln N}{N}} + \ln N\right) + \sqrt{\frac{T}{2}\ln\frac{1}{\delta}}$$

* By the tower rule: ($\mathbb{E}$ wrt decision-maker's internal randomization $U_1, U_2, \dots$)

$$\forall T, \quad \mathbb{E}[R_T] \leq C\left(\sqrt{sT\frac{\ln N}{N}} + \ln N\right).$$

We now turn to the lower bound.

Interestingly it will be proved by taking iid vectors $(\ell_{1t}, \dots \ell_{Nt})$: it will be a purely stochastic lower bound, while the upper bound is deterministic and robust.

Theorem. Given $s \in \{1..N\}$, for all strategies of the decision maker, for all $N \geq 2$ and all $T > N \ln N / 16 s$, there exists a fixed-in-advance sequence of loss vectors $(\ell_{1t}, \dots \ell_{Nt}) \in [0,1]^N$ with each at most s non-zero components, such that

$$\mathbb{E}[R_T] = \sum_{t=1}^{T} \mathbb{E}[\ell_{I_t,t}] - \min_{k=1..N} \sum_{t=1}^{T} \ell_{kt} \geq \frac{1}{32}\sqrt{sT\frac{\ln N}{N}}$$

($\mathbb{E}$ wrt the internal randomization)

Credits: Asymptotic proof technique:
Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire, Warmuth (1997)
Kwon and Perchet (2015) follows it and adapt it because of some correlations

We (Gerchinovitz, Lenard, Stoltz, 2017?) start as in Kwon and Perchet till asymptotic (central-limit-theorem) arguments are used, at which point we resort to Fano's inequality.

Proof: We fix $\varepsilon \in (0, S/2N)$.

For $i \in \{1 \dots N\}$, $\mathbb{P}_i$ is such that the random vectors $(L_{1t} \dots L_{Nt})$ are iid as t varies, with common distribution given by:

two-step randomization:
- S_t is a uniform random subset of $\{1, \dots N\}$ with S elements
- the conditional law of L_{jt} given S_t is:
  - δ_0, Dirac mass at 0 if $j \notin S_t$ (cf. sparsity constraint)
  - $\mathrm{Ber}(1/2)$ if $j \in S_t$ and $j \neq i$
  - $\mathrm{Ber}(1/2 - \varepsilon \frac{N}{S})$ if $j = i$ and $i \in S_t$

under $\mathbb{P}_i$, $\rightarrow$ L_{it} is "favored"

Marginal distributions:
$$\begin{cases} \mathbb{P}_i^{L_{jt}} = \mathrm{Ber}(S/2N) & \text{if } j \neq i \\ \mathbb{P}_i^{L_{it}} = \mathrm{Ber}(S/2N - \varepsilon) & \text{if } j = i \end{cases}$$

Thus
$$\mathbb{E}_i\left[\mathbb{E}[R_T]\right] = \mathbb{E}_i\left[\sum_{t=1}^{T} \mathbb{E}[L_{I_t,t}] - \min_{k=1\dots N} \sum_{t=1}^{T} L_{kt}\right]$$
($\mathbb{E}_i$: wrt the L_{jt}; $\mathbb{E}$: wrt the decision-maker's internal randomization $U_1, U_2, \dots$ iid $\sim \mathcal{U}[0,1]$)
$$\geq \mathbb{E}_i \otimes \mathbb{E}\left[\sum_{t=1}^{T} L_{I_t,t}\right] - \min_{k=1\dots N} \mathbb{E}_i[L_{kt}]$$
$$= \mathbb{E}_i[L_{it}] = T\left(\frac{S}{2N} - \varepsilon\right)$$

By the tower rule (applied twice):
$$\mathbb{E}_i \otimes \mathbb{E}\left[L_{I_t,t} \mid \sigma(H_{t-1}), L_{1t}, \dots L_{Nt}\right] = \sum_{j=1\dots N} \underbrace{L_{jt}}_{\text{independent of } \sigma(H_{t-1})} \mathbb{P}\{I_t = j \mid \sigma(H_{t-1})\}$$

thus (we integrate over the L_{jt})
$$\mathbb{E}_i \otimes \mathbb{E}\left[L_{I_t,t} \mid \sigma(H_{t-1})\right] = \sum_{j=1\dots N} \left(\frac{S}{2N} - \varepsilon \mathbb{1}_{\{j=i\}}\right) \times \mathbb{P}\{I_t = j \mid \sigma(H_{t-1})\}$$

and finally,
$$\mathbb{E}_i \otimes \mathbb{E}\left[L_{I_t,t}\right] = \frac{S}{2N} - \varepsilon\, \mathbb{P}_i \otimes \mathbb{P}\{I_t = i\}$$

Summing over $t = 1, \dots T$:

$$\mathbb{E}_i\left[\mathbb{E}[R_T]\right] \geq T\frac{s}{2N} - \varepsilon \sum_{t=1}^{T} \mathbb{P}_i \otimes \mathbb{P}\{I_t = i\} - \left(T(s/2N - \varepsilon)\right)$$

$$= T\varepsilon\left(1 - \mathbb{E}_i \otimes \mathbb{E}\left[\frac{N_i(T)}{T}\right]\right)$$

where $N_i(T) = \sum_{t=1}^{T} \mathbb{1}_{\{I_t = i\}}$

Now, $\underbrace{\sup_{\substack{(\ell_{it} \dots \ell_{Nt}) \in [0,1]^N \\ t=1,\dots T}} \mathbb{E}[R_T]}_{\text{our aim}} \geq \max_{i \in \{1 \dots N\}} \mathbb{E}_i\left[\mathbb{E}[R_T]\right] \geq \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_i\left[\mathbb{E}[R_T]\right]$

$$\geq T\varepsilon\left(1 - \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_i \otimes \mathbb{E}\left[\frac{N_i(T)}{T}\right]\right)$$

The main difference wrt Kwon and Perchet arises now : we replace a central-limit-theorem argument by an application of Fano's inequality.

$\mathbb{Q}$ such that $\begin{cases} S_t \text{ is a uniform random subset of } \{1 \dots N\} \text{ with } s \text{ elements} \\ \text{the conditional law of } L_{jt} \text{ given } S_t \text{ is} \\ \quad * \ \delta_0 \quad \text{if } j \notin S_t \\ \quad * \ \mathrm{Ber}(1/2) \quad \text{if } j \in S_t \end{cases}$

↗ same definition as the $\mathbb{P}_i$, except that no component is favored

Fano's lemma :

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_i \otimes \mathbb{E}\left[\frac{N_i(T)}{T}\right] \leq \frac{1}{N}\sum_{i=1}^{N} \underbrace{\mathbb{E}_{\mathbb{Q}} \otimes \mathbb{E}\left[\frac{N_i(T)}{T}\right]}_{=1} + \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N} KL(\mathbb{P}_i^T \otimes U^T, \mathbb{Q}^T \otimes U^T)}{-\ln\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbb{Q}} \otimes \mathbb{E}\left[\frac{N_i(T)}{T}\right]\right)}}$$

(We apply Fano not on the underlying probability space, but on the image space of the L_{jt} and U_t) ↪ where $\mathbb{P}_i^T = \mathbb{P}_i^{((L_{1t} \dots L_{Nt})_{t=1\dots T})}$

$$= \left(\mathbb{P}_i^{(L_{1T} \dots L_{NT})}\right)^{\otimes T}$$

and $\mathbb{Q}^T = \left(\mathbb{Q}^{(L_{1T} \dots L_{NT})}\right)^{\otimes T}$

and $U^T = \left(U_{[0,1]}\right)^{\otimes T}$

We have, by independence : $KL(\mathbb{P}_i^T \otimes U^T, \mathbb{Q}^T \otimes U^T)$

$$= T \ KL\left(\mathbb{P}_i^{(L_{1T} \dots L_{NT})}, \mathbb{Q}^{(L_{1T} \dots L_{NT})}\right)$$

we will it shows in a minute ↘ $\leq T \frac{s}{N} kl\left(\frac{1}{2} - \varepsilon N/s, 1/2\right) \leq \frac{4N}{s}\varepsilon^2 T$

Putting all elements together :

$$\frac{1}{N}\sum_{i=1}^{N} E_i \otimes E\left[\frac{N_i(T)}{T}\right] \leq \frac{1}{N} + \sqrt{\frac{4NT\varepsilon^2}{s \ln N}}$$

thus $$\sup E[R_T] \geq T\varepsilon\left(1 - \frac{1}{N}\sum_{i=1}^{N} E_i \otimes E\left[\frac{N_i(T)}{T}\right]\right)$$

$$\geq T\varepsilon\left(\underbrace{1 - \frac{1}{N}}_{\geq 1/2} - \sqrt{\frac{4NT\varepsilon^2}{s \ln N}}\right) \geq T\varepsilon\left(\tfrac{1}{2} - c\varepsilon\right)$$

Optimizing the bound (maximizing): $\varepsilon = 1/4c$ where $c = \sqrt{\frac{4NT}{s \ln N}}$ leads to $T\varepsilon/4$, which is the stated bound

But we have to check $\varepsilon < s/2N$ (cf. first line of the proof):

ok if $T > (N \ln N)/16\, s$.

So, it only remains to prove :

$$\begin{cases} KL\left(\mathbb{P}_i^{(L_{1T} \cdots L_{NT})}, \mathbb{Q}^{(L_{1T} \cdots L_{NT})}\right) \leq kl\left(\frac{1}{2} - \frac{\varepsilon N}{s}, \frac{1}{2}\right) \\ kl\left(\frac{1}{2} - \frac{\varepsilon N}{s}, \frac{1}{2}\right) \leq \frac{4N^2\varepsilon^2}{s^2} \end{cases}$$

for $u \in (0, 1/2)$, $kl(1/2 - u, 1/2) = \left(\frac{1}{2}-u\right)\ln(1-2u) + \left(\frac{1}{2}+u\right)\ln(1+2u)$

$$\leq \left(\tfrac{1}{2}-u\right)(-2u) + \left(\tfrac{1}{2}+u\right)2u = 4u^2$$

$\mathbb{P}_i^{(L_{1T} \cdots L_{NT})}$ and $\mathbb{Q}^{(L_{1T} \cdots L_{NT})}$ are convex combinations of probability distributions of the following form, indexed by the subsets of $\{1 \cdots N\}$ with s elements and up to permutations of the Bernoulli distributions in the products below (which does not change the value of the KL divergence) :

[when i is picked] $\binom{N-1}{s-1}$ distributions of the form

$$Ber\left(\tfrac{1}{2} - \frac{\varepsilon N}{s}\right) \otimes Ber\left(\tfrac{1}{2}\right)^{\otimes(s-1)} \otimes \delta_0^{\otimes(N-s)}$$

and $$Ber\left(\tfrac{1}{2}\right)^{\otimes s} \otimes \delta_0^{\otimes(N-s)}$$

[when i is not picked] $\binom{N-1}{s}$ distributions of the form

$$Ber\left(\tfrac{1}{2}\right)^{\otimes s} \otimes \delta_0^{\otimes(N-s)}$$

and $$Ber\left(\tfrac{1}{2}\right)^{\otimes s} \otimes \delta_0^{\otimes(N-s)}$$

Only the first set of distributions contributes to the Kullback-Leibler divergence:

$$KL\left(\mathbb{P}^{(L_{1T}\cdots L_{NT})}, \mathbb{Q}^{(L_{1T}\cdots L_{NT})}\right) \leq \underbrace{\frac{\binom{N-1}{S-1}}{\binom{N}{S}}}_{=S/N} KL\Big(\underbrace{Ber(1/2-\tfrac{\varepsilon N}{S}) \otimes \cancel{Ber(1/2)^{\otimes(S-1)} \otimes \delta_0}}_{Ber(1/2) \otimes \cancel{Ber(1/2)^{\otimes(S-1)} \otimes \delta_0}} \Big) = kl(1/2-\tfrac{\varepsilon N}{S}, 1/2)$$

by joint convexity of the KL