Calibration of EWA

Calibration of η for the EWA strategy.

Setting: At each round $t = 1, 2, \ldots T$,

- The statistician and the opponent simultaneously pick $p_t \in \mathcal{X}$ and $\underline{\ell}_t = (\ell_{1t}, \ldots \ell_{Nt}) \in [m, M]^N$
- p_t and $\underline{\ell}_t$ are publicly revealed

Aim: Control the regret $R_T = \sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \ell_{jt} - \min_{k=1,\ldots N} \sum_{t=1}^{T} \ell_{kt}$

Algorithm: EWA with fixed learning rate $\eta > 0$

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

Performance bound: $R_T \leq \frac{\ln N}{\eta} + \eta \frac{(M-m)^2}{8} T$

For the choice $\eta^* = \frac{1}{M-m}\sqrt{\frac{8 \ln N}{T}}$ we get $R_T \leq (M-m)\sqrt{\frac{T}{2} \ln N}$

but there are two issues with this choice:

1. T, m and M are not always known in advance
2. This η^* has often a poor performance in practice

Solutions for 1. : « doubling trick » and $p_t(\eta_t)$ where η_t varies over time

2. : $p_t(\eta_t)$ again, but with different choices for η_t

Solution for 2.?

How do we pick the η_t in practice for EWA?

What my PhD students and I do:

$$\hat{L}_t(\eta) = \sum_{s=1}^{t} \sum_{j=1}^{N} p_{j,s}(\eta)\, \ell_{j,s}$$

cumulative loss of EWA with fixed $\eta > 0$ on rounds 1 to t

Then we select $\eta_t \in \operatorname*{argmin}_{\eta>0} \hat{L}_{t-1}(\eta)$ → Exhibits good empirical performance (much better than when η tuned as prescribed by theory).

Trade off between

- not overfitting (still EWA prediction)
- fitting η according to the true stochasticity / non-stochasticity of the data (the more stochastic, the larger η; the more adversarial, the smaller η).

Open questions ×2 (would be good topics for PhD work):

1. One can show that in general, the regret of the above strategy can be linear. Can you find a calibration strategy that works well both in practice and in theory for EWA?

2. Can you quantify how to set η depending on the degree of stochasticity of the data?

Note: See how different the above calibration strategy is from, e.g., cross-validation → We really exploit here the sequential fashion of our setting!

 Lecture 2 — Winter 2022 term Lecture 2 — Winter 2022 term

Calibration of η for the EWA strategy:

The cheap way: the doubling trick.

First solution for 1

Suppose m and M are known (we'll see how to deal with m and M unknown later on). Then: for each $r = 1, 2, \ldots$

RESTART EWA with $\displaystyle \eta_r = \frac{1}{M-m}\sqrt{\frac{8 \ln N}{2^r}}$ for rounds $t = \underbrace{2^r+1, \ldots, 2^{r+1}}_{2^r \text{ such rounds}}$

(we deal with rounds $t=1$ and $t=2$ by picking uniform weights)

Fix $T \geq 3$:
Denote by r_T the smallest $r \geq 1$ such that $T \leq 2^{r+1}$ (in particular, $2^{r_T} < T$)

We have that [Global regret $\leq$ Sum of local regrets]

ie: $\displaystyle R_T = \sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \ell_{jt} - \min_k \sum_{t=1}^{T} \ell_{kt}$

max of a sum $\leq$ sum of the max.

$$\leq 2(M-m) + \sum_{r=1}^{r_T-1} \left(\sum_{t=2^r+1}^{2^{r+1}} \sum_j p_{jt}\ell_{jt} - \min_k \sum_{t=2^r+1}^{2^{r+1}} \ell_{kt} \right)$$

rounds $t=1$ and $t=2$

$$+ \underbrace{\sum_{t=2^{r_T}+1}^{T} \sum_j p_{jt}\ell_{jt} - \min_k \sum_{t=2^{r_T}+1}^{T} \ell_{kt}}$$

performance bound of EWA with η_r:

$$\leq \frac{\ln N}{\eta_r} + \eta_r (M-m)^2 \frac{2^r}{8}$$

$$= (M-m)\sqrt{\frac{2^r}{2}\ln N}$$

performance bound of EWA with η_{r_T}:

is $\displaystyle \leq \frac{\ln N}{\eta_{r_T}} + \eta_{r_T} (M-m)^2 \frac{(T-2^{r_T})}{8}$

$T \leq 2^{r_T+1}$

$$\leq \frac{\ln N}{\eta_{r_T}} + \eta_{r_T} (M-m)^2 \frac{2^{r_T}}{8}$$

$$= (M-m)\sqrt{\frac{2^{r_T}}{2}\ln N}$$

Sum of local regret bounds

We have proved:

$$\underset{\text{Global regret}}{R_T} \leq 2(M-m) + \Bigg(\underbrace{\sum_{r=1}^{r_T} \sqrt{2^r}}_{= \sqrt{2} \times \frac{(\sqrt{2})^{r_T} - 1}{\sqrt{2}-1}} \Bigg) \sqrt{\frac{\ln N}{2}} \times (M-m)$$

Since $2^{r_T} < T$, we have $(\sqrt{2})^{r_T} < \sqrt{T}$

 Lecture 2 — Winter 2022 term Lecture 2 — Winter 2022 term

The final bound reads :
$$R_T \leq \underbrace{\frac{(M-m)}{(\sqrt{2}-1)}\sqrt{T \ln N}}_{} + 2(M-m)$$

the main difference w.r.t. bound when T is known is an additional $\sqrt{2}/\sqrt{2}-1 \approx 3.42$ → measures the price for the adaptivity in T if one uses the doubling trick

(Note: Using regimes of lengths 2^r is somewhat arbitrary; lengths a^r could be considered and then a is to be optimized)

NOTE : An extension of the above argument is possible when m and M are also unknown : we take fresh starts whenever "t doubles" or whenever the current "estimates" of m and M are exceeded (and in the fresh start, their value is doubled). But it's messy to write : I don't even dare giving it to you as an exercise... because I would need to write the solution !

Here's a more elegant way to cope with the problem.

The smarter way : past-data adaptive choice of η ← Second solution for 1.

Let $p_t(\eta) = (p_{1t}(\eta), \dots, p_{Nt}(\eta))$ denote the vector of weights recommended by EWA with fixed $\eta > 0$ at round t:
$$p_{jt}(\eta) = \exp\Big(-\eta \sum_{s=1}^{t-1} \ell_{js}\Big) \Big/ \sum_{i=1}^{N} \exp\Big(-\eta \sum_{s=1}^{t-1} \ell_{is}\Big).$$

We now consider a rule to select $\eta_t > 0$ based on the past information ℓ_{js}, $s \leq t-1$, $j \in \{1..N\}$, and use the weights $p_t(\eta_t)$ at round t.
(Note that the choice of η_1 is actually irrelevant....)

Lemma : If the selected η_t are non-increasing, then for all $\ell_{kt} \in \mathbb{R}$ (not necessarily bounded),
$$R_T = \sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt}\ell_{jt} - \min_k \sum_{t=1}^{T} \ell_{kt}$$
$$\leq \frac{\ln N}{\eta_T} + \sum_{t=1}^{T} \delta_t$$
where
$$\delta_t = \sum_{j=1}^{N} p_{jt}\ell_{jt} + \frac{1}{\eta_t} \ln\Big(\sum_{j=1}^{N} p_{jt} e^{-\eta_t \ell_{jt}}\Big)$$

$\delta_t \geq 0$ by Jensen's inequality

Note : → δ_t was usually bounded via Hoeffding's inequality :
if $\ell_{jt} \in [m, M]$ $\forall j$, then $\delta_t \leq \frac{\eta_t}{8}(M-m)^2$ → but we will also consider here other bounds.

 Lecture 2 — Winter 2022 term Lecture 2 — Winter 2022 term

<u>Exercise 1</u>: If m and M are known, which η_t could you pick and which $(M-m)\sqrt{T \ln N}$ would you get? (ie, which constant would you get?).

(easy: see it as a warm-up!)

<u>Proof</u> (of the lemma):

By definition of δ_t,
$$\sum_{j=1}^{N} p_{jt}\ell_{jt} = \delta_t - \frac{1}{\eta_t}\ln\Big(\sum_j p_{jt}e^{-\eta_t \ell_{jt}}\Big)$$

The question is thus to control
$$\sum_{t=1}^{T} -\frac{1}{\eta_t}\ln\Big(\sum_j p_{jt}e^{-\eta_t\ell_{jt}}\Big) \qquad \Big[\leqslant \frac{\ln N}{\eta_T} + \min_k \sum_{t=1}^{T}\ell_{kt}\,?$$

↳ can we get some telescoping argument?

An initial transformation is useful: $\eta_{t+1} \leq \eta_t$ so that $x \mapsto x^{\eta_t/\eta_{t+1}}$ is convex:

$$\frac{1}{N}\sum_j p_{jt}e^{-\eta_t\ell_{jt}} = \frac{1}{N}\sum_j \Big(p_{jt}^{\eta_{t+1}/\eta_t} e^{-\eta_{t+1}\ell_{jt}}\Big)^{\eta_t/\eta_{t+1}}$$

$$\underset{\text{Jensen.}}{\geqslant} \Big(\frac{1}{N}\sum_j p_{jt}^{\eta_{t+1}/\eta_t}e^{-\eta_{t+1}\ell_{jt}}\Big)^{\eta_t/\eta_{t+1}}$$

$$= \frac{1}{N^{\eta_t/\eta_{t+1}}}\left(\sum_j \frac{e^{-\eta_{t+1}\sum_{s=1}^{t-1}\ell_{js}}}{\Big(\sum_k e^{-\eta_t\sum_{s=1}^{t-1}\ell_{ks}}\Big)^{\eta_{t+1}/\eta_t}}\, e^{-\eta_{t+1}\ell_{jt}}\right)^{\eta_t/\eta_{t+1}}$$

$$= \frac{1}{N^{\eta_t/\eta_{t+1}}}\,\frac{\Big(\sum_j \exp\big(-\eta_{t+1}\sum_{s=1}^{t}\ell_{js}\big)\Big)^{\eta_t/\eta_{t+1}}}{\sum_k \exp\big(-\eta_t\sum_{s=1}^{t-1}\ell_{ks}\big)}$$

Therefore, (taking $-\frac{1}{\eta_t}\ln$ in both sides)
$$-\frac{1}{\eta_t}\ln\Big(\sum_j p_{jt}e^{-\eta_t\ell_{jt}}\Big) \leq -\frac{1}{\eta_t}\ln\left(N \times N^{-\eta_t/\eta_{t+1}}\frac{\Big(\sum_j e^{-\eta_{t+1}\sum_{s=1}^{t}\ell_{js}}\Big)^{\eta_t/\eta_{t+1}}}{\sum_k e^{-\eta_t\sum_{s=1}^{t-1}\ell_{ks}}}\right)$$

Summing over $t = 1, \ldots T$:
$$\sum_{t=1}^{T} -\frac{1}{\eta_t}\ln\Big(\sum_j p_{jt}e^{-\eta_t\ell_{jt}}\Big) \leqslant \sum_{t=1}^{T}\left(-\frac{1}{\eta_t}\ln\Big(N^{1-\eta_t/\eta_{t+1}}\Big) - \frac{1}{\cancel{\eta_t}}\frac{\cancel{\eta_t}}{\eta_{t+1}}\ln\Big(\sum_j e^{-\eta_{t+1}\sum_{s=1}^{t}\ell_{js}}\Big) + \frac{1}{\eta_t}\ln\Big(\sum_k e^{-\eta_t\sum_{s=1}^{t-1}\ell_{ks}}\Big)\right)$$

(telescoping)
$$= \left(\sum_{t=1}^{T}\Big(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\Big)\ln N\right) - \frac{1}{\eta_{T+1}}\ln\Big(\sum_j e^{-\eta_{T+1}\sum_{s=1}^{T}\ell_{js}}\Big) + \underbrace{\frac{1}{\eta_1}\ln\Big(\sum_k e^{-\eta_1\times 0}\Big)}_{=\ln N/\eta_1}$$

$$= \frac{\ln N}{\eta_{T+1}} - \frac{1}{\eta_{T+1}} \ln \Bigg(\underbrace{\underbrace{\sum_j e^{-\eta_{T+1} \sum_{s=1}^{T} \ell_{js}}}_{\geq e^{-\eta_{T+1} \min_k \sum_{t=1}^{T} \ell_{kt}}} \Bigg)}_{\leq \min_{k \leq N} \sum_{t=1}^{T} \ell_{kt}}$$

! we can always assume that $\eta_{T+1} = \eta_T$ (if the rule picked another η_{T+1}, just consider for the sake of the proof $\eta'_{T+1} = \eta_T$)

⇝ Would have a better way to end up with a $\ln N / \eta_T$?

<u>Comments</u>: We did 2 things with this lemma:

1 - have η possibly depend on the past (as long as $\eta_t \searrow$ over time)

2 - consider sharper bounds than the ones obtained via Hoeffding's lemma: it suffices to focus one's attention on the δ_t.

Let's try to get sharper (second-order) bounds on the δ_t (Hoeffding's bound is a zero-order bound: something extremely uniform, that only depends on the scale $[m, M]$ and not on the concentration).

<u>Intuition</u>: $e^x \simeq 1 + x + x^2/2$ when x is small & $\ln(1+u) \approx u$ when u small

thus if η_t is small $-\frac{1}{\eta_t} \ln \left(\sum_j p_{jt} e^{-\eta_t \ell_{jt}} \right) \approx -\frac{1}{\eta_t} \ln \left(1 - \sum_j \eta_t p_{jt} \ell_{jt} + \frac{1}{2} \sum_j p_{jt} (\eta_t \ell_{jt})^2 \right)$

$$\approx \sum_j p_{jt} \ell_{jt} - \frac{\eta_t}{2} \sum_j p_{jt} \ell_{jt}^2$$

That is, $\delta_t \approx \frac{\eta_t}{2} \sum_j p_{jt} \ell_{jt}^2$

Actually, we could do a better approximation job:

 Lecture 2 — Winter 2022 term Lecture 2 — Winter 2022 term

$$\delta_t = -\frac{1}{\eta_t} \ln\Big(\sum_j p_{jt}\, e^{-\eta_t\left(\ell_{jt} - \sum_k p_{kt}\ell_{kt}\right)}\Big)$$

$$\approx -\frac{1}{\eta_t} \ln\Big(1 - \sum_j p_{jt}\Big(\eta_t\big(\ell_{jt} - \sum_k p_{kt}\ell_{kt}\big)\Big) + \frac{\eta_t^2}{2}\sum_j p_{jt}\Big(\ell_{jt} - \sum_k p_{kt}\ell_{kt}\Big)^2\Big)$$

$$\approx \frac{\eta_t}{2} \underbrace{\sum_{j=1}^{N} p_{jt}\Big(\ell_{jt} - \sum_{k=1}^{N} p_{kt}\ell_{kt}\Big)^2}_{\stackrel{\text{not.}}{=}\ v_t \ \text{(a variance-like term)}} \stackrel{\text{not.}}{=} \frac{\eta_t}{2} v_t$$

The regret bound is $\approx \dfrac{\ln N}{\eta_T} + \displaystyle\sum_{t=1}^{T} \frac{\eta_t}{2} v_t$

Choosing $\eta_t \approx \dfrac{\sqrt{\ln N}}{\sqrt{V_{t-1}}}$ where $V_{t-1} = \displaystyle\sum_{s=1}^{t-1} v_s$,

we have $$\sum \eta_t v_t \approx \sqrt{\ln N} \sum_t \frac{1}{\sqrt{V_{t-1}}}\big(V_t - V_{t-1}\big)$$

$$= \sqrt{\ln N} \sum_t \underbrace{\frac{1}{\sqrt{V_{t-1}}}\Big(\sqrt{V_t} + \sqrt{V_{t-1}}\Big)}_{\leq \text{ constant } C \text{ hopefully} \simeq 2} \underbrace{\Big(\sqrt{V_t} - \sqrt{V_{t-1}}\Big)}_{\geq 0}$$

After telescoping $\displaystyle\sum_t \eta_t v_t \lesssim C\sqrt{V_T \ln N}$

and the final regret bound is $\lesssim \dfrac{C}{2}\sqrt{V_T \ln N}$

↳ For a decade (2005 - 2014) this proof was somewhat messy and suboptimal, and led to a non-homogeneous bound.
Recently, in 2014, de Rooij, van Erven, Grünwald & Koolen could at last make it 100% precise and homogeneous.

Their key argument to link δ_t and η_t, v_t is Bernstein's inequality.

<u>Lemma</u> [Bernstein's inequality] : X random variable taking values in $[m,M]$

Then: $\forall \eta > 0$, $\ln \mathbb{E}[e^{\eta X}] \leqslant \eta \mathbb{E}[X] + \left(\frac{e^{\eta(M-m)} - 1 - \eta(M-m)}{(M-m)^2} \right) \mathrm{Var}(X)$

<u>Proof</u>: $\varphi : x \in \mathbb{R} \mapsto (e^x - x - 1)/x^2$ is increasing over $\mathbb{R}$

$\eta > 0$, $X \leqslant M$ and $\mathbb{E}X \geqslant m$ so that $\eta(X - \mathbb{E}X) \leqslant \eta(M-m)$

then $\varphi(\eta(X - \mathbb{E}X)) \leqslant \varphi(\eta(M-m))$

That is, $e^{\eta(X-\mathbb{E}X)} - \eta(X - \mathbb{E}X) - 1 \leqslant \cancel{\eta^2}(X - \mathbb{E}X)^2 \, \frac{e^{\eta(M-m)} - \eta(M-m) - 1}{\cancel{\eta^2}(M-m)^2}$

Taking expectations: $\mathbb{E}[e^{\eta X}] \times e^{-\eta \mathbb{E}X} \leqslant 1 + \frac{e^{\eta(M-m)} - \eta(M-m) - 1}{(M-m)^2} \mathrm{Var}(X)$

The proof is concluded by taking logarithms and using $\ln(1+u) \leqslant u$ $\forall u > -1$.

<u>Algorithm</u>: EWA with at each round $t \geqslant 1$ $\eta_t = \frac{\ln N}{\sum_{s=1}^{t-1} \delta_s}$ with the convention that $\frac{\ln N}{0} = +\infty$

<u>Note</u>: $p_1 = (1/N, \dots 1/N)$ and $\eta_1 = +\infty$; since the value of η_1 is irrelevant this merely means that $\eta_1 \geqslant \eta_2$ and

$$\delta_1 = \frac{1}{N}\sum_j \ell_{j1} + \frac{1}{\eta_1} \ln\left(\sum_j \frac{1}{N} e^{-\eta_1 \ell_{j1}}\right)$$
$$= \frac{1}{N}\sum_j \ell_{j1} - \min_k \ell_{k1}$$

$\delta_1 > 0$ except if all losses are equal, in which case NO REGRET is suffered and the indexation of time could well start at 2...

With no loss of generality we will assume $\delta_1 > 0$.

<u>Theorem</u>: The above strategy ensures that:

$\forall m \leqslant M$, $\forall \ell_{jt} \in [m,M]$, $R_T = \sum_{t,j} p_{jt}\ell_{jt} - \min_k \sum_{t=1}^T \ell_{kt}$

not known by the algorithm

$$\leqslant 2 \sqrt{\sum_{t=1}^T \sum_j p_{jt}\left(\ell_{jt} - \sum_k p_{kt}\ell_{kt}\right)^2 \ln N}$$
$$+ (M-m)\left(2 + \frac{4}{3}\ln N\right)$$

 Lecture 2 — Winter 2022 term Lecture 2 — Winter 2022 term

Proof: $\eta_t \searrow$ thus $R_T \leq \dfrac{\ln N}{\eta_T} + \displaystyle\sum_{t=1}^{T} \delta_t$

Part 1

By Bernstein's inequality (for $t \geq 2$ so that $\eta_t < +\infty$):

$$\delta_t \leq \boxed{\frac{1}{\eta_t} \, \frac{e^{\eta_t (M-m)} - \eta_t (M-m) - 1}{(M-m)^2}} \, v_t \qquad \text{for } t \geq 2$$

(with $X_t = -\ell_{jt}$ with probability p_{jt} and $\eta_t > 0$)

$X_t \in [-M, -m]$

this factor is still $\approx \eta_t / 2$

reminder: $v_t = \sum_j \Big(\ell_{jt} - \sum_k p_{kt} \ell_{kt} \Big)^2 p_{jt}$

We make the $\delta_t \lesssim \dfrac{\eta_t v_t}{2}$ more precise, by proving $v_t \gtrsim \dfrac{2\delta_t}{\eta_t}$:

$$v_t \geq \frac{\eta_t (M-m)}{e^{\eta_t (M-m)} - \eta_t (M-m) - 1} (M-m) \delta_t \qquad \text{for } t \geq 2$$

$$= \underbrace{\frac{2}{\eta_t (M-m)} (M-m) \delta_t}_{\text{what we want}} - \underbrace{(M-m) \delta_t \times f\big(\eta_t (M-m)\big)}_{\text{the gap between what we want and what we have}}$$

where for $x \geq 0$: $f(x) = \dfrac{2}{x} - \dfrac{x}{e^x - x - 1}$

One can show that f is decreasing, so that $f(x) \leq f(0) = 2/3 \quad \forall x \geq 0$

Therefore, $v_t \geq \dfrac{2\delta_t}{\eta_t} - \dfrac{2}{3}(M-m)\delta_t$ for $t \geq 2$, but also valid for $t=1$ as $\eta_1 \to +\infty$

Part 2

$$\Big(\sum_{t=1}^{T} \delta_t \Big)^2 \underset{\uparrow \text{ telescoping}}{=} \sum_{t=1}^{T} \bigg(\Big(\sum_{s=1}^{t} \delta_s \Big)^2 - \Big(\sum_{s=1}^{t-1} \delta_s \Big)^2 \bigg)$$

$$\underset{\substack{a^2 - b^2 \\ = (a+b)(a-b)}}{=} \sum_{t=1}^{T} \delta_t \Big(\underbrace{\delta_t}_{\substack{\text{direct bound:} \\ \text{this is } \leq M-m}} + 2 \underbrace{\sum_{s=1}^{t-1} \delta_s}_{\substack{\text{by def of the algorithm, this} \\ = \frac{\ln N}{\eta_t}, \text{ even for } t=1}} \Big)$$

$$\leq \sum_{t=1}^{T} \Big((M-m)\delta_t + \underbrace{\frac{2\delta_t}{\eta_t} \ln N}_{\substack{\text{by Part 1:} \\ \leq v_t \ln N + \frac{2}{3}(M-m)\delta_t \ln N}} \Big)$$

Summarizing:

$$\Big(\sum_{t=1}^{T} \delta_t \Big)^2 \leq V_T \ln N + (M-m)\Big(1 + \frac{2}{3} \ln N\Big) \sum_{t=1}^{T} \delta_t$$

reminder: $V_T = \sum_{t \leq T} v_t$

This is a 2nd-order inequation: $x^2 \leq b + ax$ for $a, b \geq 0$ $\Rightarrow$ $x \leq a + \sqrt{b}$

[indeed, x is then smaller than the larger root of the 2nd-order polynom $u \mapsto u^2 - au - b$, that is,

$$x \leqslant \frac{a + \sqrt{a^2+4b}}{2} \leqslant a + \sqrt{b} \quad \text{using} \quad \sqrt{t+t'} \leqslant \sqrt{t} + \sqrt{t'}. \;]$$

We thus get :
$$\sum_{t=1}^{T} \delta_t \leqslant (M-m)\left(1+\frac{2}{3}\ln N\right) + \sqrt{V_T \ln N}$$

On the other hand:
$$\frac{\ln N}{\eta_T} \leqslant \frac{\ln N}{\eta_{T+1}} = \sum_{t=1}^{T} \delta_t$$

Final bound :
$$R_T \leqslant \frac{\ln N}{\eta_T} + \sum_{t=1}^{T} \delta_t \leqslant 2 \sum_{t=1}^{T} \delta_t \leqslant 2 \times \text{this bound}$$

<u>Exercise 2.</u> The second part of the proof (as well as the definition of the η_t) seems unnatural to me. Can you get a more natural analysis ?

<u>Corollary</u>: The algorithm above is such that

$\forall T$, $\forall m \leqslant M$, $\forall \ell_{j,t} \in [m, M]$,
$$\sum_{t,j} p_{j,t}\,\ell_{j,t} - \min_k \sum_{t=1}^{T} \ell_{k,t} \leqslant (M-m)\sqrt{T \ln N} + (M-m)\left(2 + \frac{4}{3}\ln N\right)$$

not known by the algorithm

$\hookrightarrow$ The algorithm is adaptive to T and to the bounds m and M.

<u>Proof</u>: v_t = some variance of a random variable $\in [m, M]$ $\leqslant \dfrac{(M-m)^2}{4}$, as in the proof of Hoeffding's lemma.

<u>Exercise 3.</u> Improvement for small losses.

This exercise shows that the $\sqrt{T}$ order of magnitude for the regret can be improved on easy problems. The results below are of a nature similar to the "fast rates" results in classification / statiscal

Lecture 2 — Winter 2022 term Lecture 2 — Winter 2022 term

learning.

(1) Assume that losses ℓ_{jt} are non-negative: $\ell_{jt} \geq 0$, i.e., $\ell_{jt} \in [0, M]$.

Show that $\displaystyle\sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \Big(\ell_{jt} - \sum_k p_{kt} \ell_{kt} \Big)^2 \leq M \sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \ell_{jt}$.

(2) Exhibit a regret bound of order $\displaystyle\sqrt{M \min_{k=1\ldots N} \sum_{t=1}^{T} \ell_{kt} \ln N}$

(3) Specify the value of this bound when $\exists k \mid \displaystyle\sum_{t=1}^{T} \ell_{kt} = 0$

(the « perfect expert » case)

↳ Have the bound be really reader-friendly, please!

<u>Exercise 4.</u> « Impossible tuning »

Consider the PROD forecaster: $p_1 = (1/N, \ldots 1/N)$

and for $t \geq 2$, $\displaystyle p_{jt} = \prod_{s=1}^{t-1} (1 - \eta \ell_{js}) \Big/ \sum_{k=1}^{N} \prod_{s=1}^{t-1} (1 - \eta \ell_{ks})$

It appears as a first-order approximation to EWA.

(1) Show that $\displaystyle\sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \ell_{jt} - \sum_{t=1}^{T} \ell_{kt} \leq \frac{\ln N}{\eta} + \eta \sum_{t=1}^{T} \ell_{kt}^2$

when losses are such that $\ell_{jt} \leq M$ and when η has been chosen so that $\eta \leq 1/2M$.

Hint: use $\ln(1+u) \geq u - u^2 \quad \forall u \geq -1/2$ and $\ln(1+u) \leq u \quad \forall u > -1$

(2) Can you tune η (by doubling trick, by having it depend on time, etc.) so that:

$$\sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \ell_{jt} \leq \min_{k=1\ldots N} \left\{ \sum_{t=1}^{T} \ell_{kt} + C \sqrt{\sum_{t=1}^{T} \ell_{kt}^2 \ln N} \right\}$$

(for some universal constant C to be defined by the analysis) ?

IF SO: YOU'LL GET 20/20 AT MY EXAM AND WE WILL WRITE A PAPER TOGETHER!

↳ a paper by Grünwald / Koolen / van Erven gets such a bound but for a different algorithm.

The Hoeffding–Azuma inequality

The Hoeffding-Azuma inequality

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 (ie, $\forall t \geq 1$, X_t is $\mathcal{F}_t$-measurable), that are bounded: $\forall t$, $a_t \leqslant X_t \leqslant b_t$ a.s., where $a_t, b_t \in \mathbb{R}$.

Then (« probabilistic version »)

$$\forall \varepsilon > 0, \qquad \mathbb{P}\left\{ \sum_{t=1}^{T} X_t - \sum_{t=1}^{T} \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \geqslant \varepsilon \right\} \leqslant \exp\left(- \frac{2\varepsilon^2}{\sum_{t=1}^{T}(b_t - a_t)^2} \right)$$

or (« statistical version », totally equivalent)

$\forall \delta \in (0,1)$, with probability at least $1-\delta$,

$$\sum_{t=1}^{T} X_t - \sum_{t=1}^{T} \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \leqslant \sqrt{\frac{\sum_{t=1}^{T}(b_t-a_t)^2}{2} \ln \frac{1}{\delta}}$$

Note: Hoeffding's inequality is the special case when all X_t are independent and $\mathcal{F}_{t-1} = \sigma(X_1 \ldots X_{t-1})$, so that $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[X_t]$.

Basic ingredient of the proof: extension of Hoeffding's lemma to conditional expectations

Lemma: X random variable s.t. $X \in [a,b]$ a.s.

Then, for all σ-algebras $\mathcal{G}$, for all $s \in \mathbb{R}$,

$$\ln \mathbb{E}\left[e^{s(X - \mathbb{E}[X \mid \mathcal{G}])} \,\middle|\, \mathcal{G} \right] = \ln\left(\mathbb{E}\left[e^{sX} \mid \mathcal{G} \right] \right) - s\,\mathbb{E}[X \mid \mathcal{G}] \leqslant \frac{s^2}{8}(b-a)^2$$

(We will discuss the proof later on... let's first prove the theorem based on this lemma.)

Proof (of the theorem):

Markov-Chernov bounding (= Markov's inequality after taking exponents):

We denote
$$\underbrace{S_T}_{\text{(martingale}} = \underbrace{\sum_{t=1}^{T}}_{= \text{ sum of}} \underbrace{X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]}_{\substack{\text{martingale increments)} \\ \text{or martingale differences}}}$$

The « probabilistic version » is about upper bounding $\mathbb{P}\{S_T > \varepsilon\}$:

$$\mathbb{P}\{S_T > \varepsilon\} \underset{\forall s>0}{=} \mathbb{P}\{e^{sS_T} > e^{s\varepsilon}\} \underset{\text{Markov's inequality}}{\leqslant} e^{-s\varepsilon}\,\mathbb{E}[e^{sS_T}]$$

We show by induction that $\mathbb{E}[e^{sS_T}] \leqslant \exp\left(\frac{s^2}{8}\sum_{t=1}^{T}(b_t-a_t)^2\right)$

- For $T=1$, true by the conditional version of Hoeffding's lemma and the fact that $S_1 = X_1 - \mathbb{E}[X_1 \mid \mathcal{F}_0]$ with $X_1 \in [a_1, b_1]$
  + taking expectations by tower rule: $\mathbb{E} = \mathbb{E}[\mathbb{E}[\ \mid \mathcal{F}_0]]$
- For $T-1 \to T$, where $T \geqslant 2$:

  The extension of Hoeffding's lemma ensures that
  $$\mathbb{E}\left[e^{s(X_T - \mathbb{E}[X_T \mid \mathcal{F}_{T-1}])} \,\middle|\, \mathcal{F}_{T-1}\right] \leqslant e^{s^2(b_T-a_T)^2/8}$$

  so that
  $$\begin{aligned}\mathbb{E}[e^{sS_T}] &= \mathbb{E}\left[\mathbb{E}[e^{sS_T} \mid \mathcal{F}_{T-1}]\right] \\ &= \mathbb{E}\left[e^{sS_{T-1}}\,\mathbb{E}\left[e^{s(X_T - \mathbb{E}[X_T\mid\mathcal{F}_{T-1}])} \,\middle|\, \mathcal{F}_{T-1}\right]\right] \\ &\leqslant e^{s^2(b_T-a_T)^2/8} \times \mathbb{E}[e^{sS_{T-1}}] \\ &\leqslant \exp\Big(s^2 \sum_{t\leqslant T}(b_t-a_t)^2/8\Big)\end{aligned}$$
  by the induction hypothesis

Substituting above:
$$\mathbb{P}\{S_T > \varepsilon\} \leqslant \inf_{s>0} \exp\left(-s\varepsilon + s^2\sum_{t\leqslant T}\frac{(b_t-a_t)^2}{8}\right) = \exp\left(-2\varepsilon^2 \Big/ \sum_{t\leqslant T}(b_t-a_t)^2\right).$$

strictly convex function to minimize in the exponent: minimum achieved at s^* such that $s^*\sum_{t\leqslant T}(b_t-a_t)^2/4 = \varepsilon$ (gradient vanishes)

i.e. $s^* = 4\varepsilon \Big/ \sum_{t\leqslant T}(b_t-a_t)^2$

→ It only remains to prove the extension of Hoeffding's lemma to conditional expectations.

But first (reminder) unconditional version

Lemma (Hoeffding): X random variable s.t. $X \in [a,b]$ a.s.

Then $\forall s \in \mathbb{R}$,

$$\ln \mathbb{E}\left[e^{s(X-\mathbb{E}X)}\right] = \ln \mathbb{E}\left[e^{sX}\right] - s\,\mathbb{E}[X] \leqslant \frac{s^2}{8}(b-a)^2$$

Proof (most elegant one I know of):

$$\Psi(s) = \ln \mathbb{E}\left[e^{sX}\right] \quad \text{defined for all } s \in \mathbb{R}$$

Ψ is differentiable at each $s \in \mathbb{R}$: cf. X bounded, thus $\eta \mapsto X e^{\eta X}$ locally dominated around s by an integrable r.v. independent of η thus $\eta \mapsto \mathbb{E}[e^{\eta X}]$ differentiable at s with derivative $\mathbb{E}[X e^{sX}]$

with

$$\Psi'(s) = \frac{\mathbb{E}[X e^{sX}]}{\mathbb{E}[e^{sX}]}$$

Similarly, Ψ is twice differentiable at each $s \in \mathbb{R}$, with:

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

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

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

$X \in [a,b]$:

$$\mathrm{Var}_{\mathbb{Q}}(X) = \inf_{\mu \in \mathbb{R}} \mathbb{E}_{\mathbb{Q}}\left[(X-\mu)^2\right] \leqslant \mathbb{E}_{\mathbb{Q}}\left[\left(X - \frac{a+b}{2}\right)^2\right] \leq \frac{(b-a)^2}{4}$$

Taylor: $\exists x$ s.t.

$$\Psi(s) = \underbrace{\Psi(0)}_{=0} + s\,\underbrace{\Psi'(0)}_{=\mathbb{E}[X]} + \frac{s^2}{2}\,\underbrace{\Psi''(x)}_{\leqslant (b-a)^2/4}$$

Cf. Ψ is actually C² smooth

i.e.,

$$\ln \mathbb{E}\left[e^{sX}\right] \leqslant s\,\mathbb{E}[X] + \frac{s^2}{8}(b-a)^2$$

Back to Hoeffding's lemma with conditional expectations:

Proof 1? Can we take the proof of Hoeffding's lemma we just saw and replace all $\mathbb{E}$ by $\mathbb{E}[\ |\mathcal{G}_j]$?

$\Psi(s) = \ln \mathbb{E}[e^{sX} | \mathcal{G}_j]$ → The theorem of differentiation under $\mathbb{E}[\ ]$ only requires dominated convergence, which holds true for $\mathbb{E}[\ |\mathcal{G}_j]$ as well. Thus we also have a theorem of differentiation under $\mathbb{E}[\ |\mathcal{G}_j]$:

a.s., $\Psi''(s)$ exists and equals
$$\Psi''(s) = \frac{\mathbb{E}[X^2 e^{sX} | \mathcal{G}_j]\, \mathbb{E}[e^{sX}|\mathcal{G}_j] - \left(\mathbb{E}[X e^{sX}|\mathcal{G}_j]\right)^2}{\left(\mathbb{E}[e^{sX}|\mathcal{G}_j]\right)^2}$$

= some conditional variance under a different probability measure?

Yes, see details in some pages.

However, there are two other proofs that I find more elementary:

Proof 2 Too bad for elegance, let's get back to the original proof of Hoeffding's (unconditional) lemma, which only relies on calculus:

$Y = X - \mathbb{E}[X|\mathcal{G}_j] \in [A, B]$ where $A = a - \mathbb{E}[X|\mathcal{G}_j] \leqslant 0$, $B = b - \mathbb{E}[X|\mathcal{G}_j] \geqslant 0$ are both $\mathcal{G}_j$-measurable and $B - A = b - a > 0$

$$Y = \underbrace{\frac{B-Y}{B-A}}_{\text{convex weights}} A + \underbrace{\frac{Y-A}{B-A}}_{} B$$

Since $y \mapsto e^{sy}$ is convex:

$$e^{sY} \leqslant \frac{B-Y}{B-A} e^{sA} + \frac{Y-A}{B-A} e^{sB}$$

Taking $\mathbb{E}[\ |\mathcal{G}_j]$: using $\mathbb{E}[Y|\mathcal{G}_j] = 0$ and A, B $\mathcal{G}_j$-measurable:

$$\mathbb{E}[e^{sY}|\mathcal{G}_j] \leqslant \frac{B}{B-A} e^{sA} - \frac{A}{B-A} e^{sB}$$

note that $B/B-A$ and $-A/B-A$ are convex weights

Now, by a function study (the very same as the one we performed in the proof of the unconditional version of Hoeffding's lemma) — or even by the latter lemma itself:

$\forall u, v \in \mathbb{R}, \ \forall p \in [0,1], \ \forall s \in \mathbb{R},$
$$\ln\left(p e^{su} + (1-p)e^{sv}\right) \leqslant s\big(pu + (1-p)v\big) + \frac{s^2}{8}(v-u)^2$$

← ln of expected value of e^{sZ} where $Z = \begin{cases} u & \text{w.p. } p \\ v & 1-p \end{cases}$

expected value of Z

range is $[u,v]$

In particular,
$$\frac{B}{B-A} e^{sA} - \frac{A}{B-A} e^{sB} \leqslant \exp\left(s\left(\frac{BA}{B-A} - \frac{AB}{B-A}\right) + \frac{s^2}{8}(B-A)^2 \right)$$
$$= \exp\left(\frac{s^2}{8}(b-a)^2\right)$$
recall that a.s., $B-A = b-a$

Summarizing:
$$\mathbb{E}\left[e^{sY} \mid \mathcal{G}\right] \leqslant \exp\left(\frac{s^2}{8}(b-a)^2\right)$$
$$= \mathbb{E}\left[e^{sX} \mid \mathcal{G}\right] \times \exp\left(-s\,\mathbb{E}[X \mid \mathcal{G}]\right)$$

<u>Proof 3</u> My preferred (not only because I found it by myself): Hoeffding's lemma in its unconditional version <u>ENTAILS</u> the conditional version! This is because Hoeffding's lemma holds for all probability distributions — we should play with this fact.

For all $A \in \mathcal{G}$ s.t. $\mathbb{P}(A) > 0$, let $\mathbb{P}_A = \mathbb{P}(\,\cdot \mid A)$, the conditional distribution given the event A.

The unconditional version of Hoeffding's lemma ensures that
$$\forall A \in \mathcal{G} \text{ s.t. } \mathbb{P}(A) > 0, \quad \forall s \in \mathbb{R}, \qquad \ln \mathbb{E}_A\left[e^{sX}\right] \leqslant s\,\mathbb{E}_A[X] + \frac{s^2}{8}(b-a)^2$$

Why do we consider the $\mathbb{E}_A$? Because $\mathbb{E}[X \mid \mathcal{G}]$ is the unique $\mathcal{G}$-measurable random variable such that $\forall A \in \mathcal{G}, \quad \mathbb{E}[X \mathbb{1}_A] = \mathbb{E}\big[\mathbb{E}[X \mid \mathcal{G}]\, \mathbb{1}_A\big]$

or, equivalently, $\forall A \in \mathcal{G}$ s.t. $\mathbb{P}(A) > 0$, $\quad \mathbb{E}_A[X] = \mathbb{E}_A\big[\mathbb{E}[X \mid \mathcal{G}]\big]$.

Now, consider the random variable $H = e^{s\,\mathbb{E}[X \mid \mathcal{G}]}\, e^{s^2(b-a)^2/8} - \mathbb{E}\left[e^{sX} \mid \mathcal{G}\right]$

We want to prove that $H \geq 0$ a.s.

H is $\mathcal{G}_j$-measurable & thus suffices to show that for all $A \in \mathcal{G}_j$ with $\mathbb{P}(A) > 0$,

$$\mathbb{E}_A[H] \geq 0 \qquad \text{that is,} \qquad \mathbb{E}[H \mathbb{1}_A] \geq 0$$

Indeed,
$$\mathbb{E}_A[H] = e^{s^2(b-a)^2/8}\, \mathbb{E}_A\left[e^{s\mathbb{E}[X|\mathcal{G}_j]}\right] - \mathbb{E}_A\left[\mathbb{E}[e^{sX}|\mathcal{G}_j]\right]$$
$$= e^{s^2(b-a)^2/8}\, \mathbb{E}_A\left[e^{s\mathbb{E}[X|\mathcal{G}_j]}\right] - \mathbb{E}_A\left[e^{sX}\right]$$
$$\underset{\text{(Jensen)}}{\geq} e^{s^2(b-a)^2/8}\, e^{s\mathbb{E}_A[X]} - \mathbb{E}_A\left[e^{sX}\right] \geq 0$$

(unconditional version of Hoeffding's lemma)

<u>Proof 1</u> Let's get back to it. In what follows all expectations relative to the original probability distribution $\mathbb{P}$ will be denoted by $\mathbb{E}$, and expectations under alternative distributions $\mathbb{Q}$ will be denoted by $\mathbb{E}_{\mathbb{Q}}$.

(1) Consider the random variable $L_s = \dfrac{e^{sX}}{\mathbb{E}[e^{sX}|\mathcal{G}_j]} \geq 0$

Since $\mathbb{E}[L_s] = \mathbb{E}\big[\mathbb{E}[L_s|\mathcal{G}_j]\big] = 1$, L_s is a density

We define the probability $\mathbb{Q}_s$ as : $\dfrac{d\mathbb{Q}_s}{d\mathbb{P}} = L_s$

(2) We show that $\varphi'(s) = \dfrac{\mathbb{E}[Xe^{sX}|\mathcal{G}_j]}{\mathbb{E}[e^{sX}|\mathcal{G}_j]}$ also equals $\mathbb{E}_{\mathbb{Q}_s}[X|\mathcal{G}_j]$

$\varphi''(s) =$ [see expression some pages ago] $\mathrm{Var}_{\mathbb{Q}_s}(X|\mathcal{G}_j)$

To do so, it suffices to prove that for all bounded random variables Z (we'll pick $Z = X$ and $Z = X^2$), we have :

$$\mathbb{E}[ZL_s|\mathcal{G}_j] = \mathbb{E}_{\mathbb{Q}_s}[Z|\mathcal{G}_j]$$

or equivalently, that $\forall A \in \mathcal{G}_j$, $\mathbb{E}\big[\mathbb{E}[ZL_s|\mathcal{G}_j]\,\mathbb{1}_A\big] = \mathbb{E}\big[\mathbb{E}_{\mathbb{Q}_s}[Z|\mathcal{G}_j]\,\mathbb{1}_A\big]$

(since both sides are $\mathcal{G}_j$-measurable)

But :
$$\mathbb{E}\big[\mathbb{E}[ZL_s|\mathcal{G}_j]\,\mathbb{1}_A\big] \underset{\text{one characterization of } \mathbb{E}[\ |\mathcal{G}_j]}{=} \mathbb{E}[ZL_s \mathbb{1}_A] \underset{L_s \text{ is } \frac{d\mathbb{Q}_s}{d\mathbb{P}}}{=} \mathbb{E}_{\mathbb{Q}_s}[Z\mathbb{1}_A]$$

 Lecture 2 — Winter 2022 term Lecture 2 — Winter 2022 term

and on the other end

$$\mathbb{E}\Big[\mathbb{E}_{Q_s}[Z|\mathcal{G}]\,\mathbb{1}_A\Big]$$

as $\mathbb{E}[L_s|\mathcal{G}] = 1$ as by definition of L_s

$$= \mathbb{E}\Big[\mathbb{E}_{Q_s}[Z|\mathcal{G}]\,\mathbb{E}[L_s|\mathcal{G}]\,\mathbb{1}_A\Big]$$

$\mathbb{E}_{Q_s}[Z|\mathcal{G}]$ and $\mathbb{1}_A$ $\mathcal{G}$-measurable and can go inside $\mathbb{E}[\;|\mathcal{G}]$

$$= \mathbb{E}\Big[\mathbb{E}\big[L_s\,\mathbb{E}_{Q_s}[Z|\mathcal{G}]\,\mathbb{1}_A\,\big|\,\mathcal{G}\big]\Big]$$

"tower rule"

$$= \mathbb{E}\Big[L_s\,\mathbb{E}_{Q_s}[Z|\mathcal{G}]\,\mathbb{1}_A\Big]$$

$L_s = \dfrac{dQ_s}{d\mathbb{P}}$

$$= \mathbb{E}_{Q_s}\Big[\mathbb{E}_{Q_s}[Z|\mathcal{G}]\,\mathbb{1}_A\Big]$$

by a characterization of $\mathbb{E}[\;|\mathcal{G}]$

$$= \mathbb{E}_{Q_s}[Z\mathbb{1}_A]$$

which concludes the proof of (2).

That's the real trick, but it's not a trick, it's what is called Bayes' formula for conditional expectations

(3) $\psi'(0) = \mathbb{E}[X|\mathcal{G}]$ (clear)

and for all x, $\psi''(x) = \mathrm{Var}_{Q_x}(X|\mathcal{G}) \leqslant \dfrac{(b-a)^2}{4}$ (we prove that below)

so that we may conclude as in the case of the unconditional Hoeffding's lemma.

Indeed, $\forall c \in \mathbb{R}$,

$$\mathbb{E}_{Q_x}\big[(X-c)^2\,|\,\mathcal{G}\big] = \mathbb{E}_{Q_x}\Big[\big(X - \mathbb{E}_{Q_x}[X|\mathcal{G}] + \mathbb{E}_{Q_x}[X|\mathcal{G}] - c\big)^2\,\Big|\,\mathcal{G}\Big]$$

$$= \underbrace{\mathbb{E}_{Q_x}\Big[\big(X - \mathbb{E}_{Q_x}[X|\mathcal{G}]\big)^2\,\Big|\,\mathcal{G}\Big]}_{\overset{\text{def.}}{=}\ \mathrm{Var}_{Q_x}(X|\mathcal{G})} + 2\times 0 + \underbrace{\cdots}_{\geqslant 0}$$

and we take

$c = \dfrac{b+a}{2}$ and use $X \in [a,b]$

to get the a.s. bound $(X-c)^2 \leq \dfrac{(b-a)^2}{4}$

 Lecture 2 — Winter 2022 term Lecture 2 — Winter 2022 term

A final remark: Dealing with non-constant but predictable ranges in the Hoeffding–Azuma inequality

↳ Sometimes useful to get slightly better constants

Hoeffding's lemma: extension #1

Setting: X random variable s.t. there exists a bounded and $\mathcal{G}$-measurable random variable G, as well as $a, b \in \mathbb{R}$ with: $G+a \leqslant X \leqslant G+b$

Then:

$$\forall s \in \mathbb{R}, \qquad \ln \mathbb{E}\left[e^{sX} \,\middle|\, \mathcal{G}\right] \leqslant s\, \mathbb{E}[X \mid \mathcal{G}] + \frac{s^2}{8}(b-a)^2$$

Rk: X bounded as well, we get this statement from the first statement by considering $a \leqslant X - G \leqslant b$

extension #2

Setting: What about when there exist U, V two $\mathcal{G}$-measurable random variables with $U \leqslant X \leqslant V$ and X bounded (for e^{sX} to be $\mathbb{L}^1$)?

An inspection of Proof 2 reveals that one can prove

$$\forall s \in \mathbb{R}, \qquad \ln \mathbb{E}\left[e^{sX} \,\middle|\, \mathcal{G}\right] \leqslant s\, \mathbb{E}[X \mid \mathcal{G}] + \frac{s^2}{8}(V-U)^2$$

Note: To state our extension of the Hoeffding–Azuma inequality, we will need a constant bound on $V-U$:

$$V-U \leqslant \Delta \qquad \text{where } \Delta \in \mathbb{R}^+$$

But actually $\left\{ \begin{array}{l} U \leqslant X \leqslant V \\ V-U \leqslant \Delta \in \mathbb{R}^+ \end{array} \right.$ entail $\dfrac{U+V}{2} - \dfrac{\Delta}{2} \leqslant X \leqslant \dfrac{U+V}{2} + \dfrac{\Delta}{2}$

So we're back to extension #1 (in particular, $(U+V)/2$ is bounded)

Hoeffding–Azuma inequality

Let $(\mathcal{F}_t)$ be a filtration and (X_t) be a sequence of adapted random variables such that

(1) $\forall t$, $\left.\begin{array}{l} \exists\, G_t \ \mathcal{F}_{t-1}\text{-measurable and bounded} \\ \exists\, a_t, b_t \in \mathbb{R} \end{array}\right\}$ with $G_t + a_t \leqslant X_t \leqslant G_t + b_t$

possibly following from

(2) $\forall t$, $\left.\begin{array}{l} \exists\, U_t, V_t \ \mathcal{F}_{t-1}\text{-measurable} \\ X_t \text{ bounded} \\ \exists\, \Delta_t \in \mathbb{R}^+ \end{array}\right\}$ with $\left\{\begin{array}{l} V_t - U_t \leqslant \Delta_t \\ U_t \leqslant X_t \leqslant V_t \end{array}\right.$

Lecture 2 — Winter 2022 term
Lecture 2 — Winter 2022 term

Then $\forall \delta \in (0,1)$

with probability at least $1-\delta$,

Case (1) $$\sum_{t=1}^{T} X_t - \sum_{t=1}^{T} \mathbb{E}\left[X_t \mid \mathcal{F}_{t-1}\right] \leqslant \sqrt{\frac{\sum_{t=1}^{T}(b_t-a_t)^2}{2} \ln \frac{1}{\delta}}$$

Case (2) $$\sum_{t=1}^{T} X_t - \sum_{t=1}^{T} \mathbb{E}\left[X_t \mid \mathcal{F}_{t-1}\right] \leqslant \sqrt{\frac{\sum_{t=1}^{T}\Delta_t^2}{2} \ln 1/\delta}$$

Randomized prediction

What can we do when no convexity assumption holds?

↳ Non-convex aggregation via randomization.

Example 1: N-ary decisions in a game (4-ary if we have to pick paths in a graph: → ← ↑ ↓) (binary if accept/reject actions)

1. Opponent picks state of the world y_t } simultaneously
2. Statistician picks action $j_t \in \{1, \dots N\}$
3. Loss $\ell(j_t, y_t)$ or reward $-\ell(j_t, y_t)$ is encountered, both y_t and j_t are made public

Example 2: Prediction with expert advice (the «meta-statistical» framework)

↳ when the prediction space is not convex:

1. Opponent picks observation $y_t \in \mathcal{Y}$
2. Simultaneously, experts provide forecasts $f_{j,t} \in \mathcal{Y}$, $j \in \{1 \dots N\}$ and statistician picks forecast $\hat{y}_t \in \mathcal{Y}$
3. y_t and $\hat{y}_t$ are revealed, losses $\ell(\hat{y}_t, y_t)$ and $\ell(f_{j,t}, y_t)$ are suffered

No convexity: $\mathcal{Y}$ not convex
eg, $\mathcal{Y} = \{1 \dots M\}$ in M-ary classification

↳ $\hat{y}_t$ cannot be any convex/linear prediction of the $f_{j,t}$ we wish.

Solution: (at least, an easy solution, there might be others)

Draw $J_t \in \{1, \dots N\}$ at random and pick
- action J_t (in Example 1)
- forecast $\hat{y}_t = f_{J_t, t}$ (in Example 2)

⇓

General setting: Simultaneously
1. Opponent picks $\underline{\ell}_t = (\ell_{1t}, \dots \ell_{Nt}) \in \mathbb{R}^N$
2. Statistician draws $J_t \in \{1, \dots N\}$

3. J_t and $(\ell_{1t}, \dots \ell_{Nt})$ are revealed.

Aim: Minimize the regret $\displaystyle\sum_{t=1}^T \ell_{J_t, t} - \min_{k=1,\dots N} \sum_{t=1}^T \ell_{kt}$

! The losses ℓ_{jt} may depend on the past, ie, on $J_1, \dots J_{t-1}$

Methodology: We denote by $p_t = (p_{1t} \cdots p_{Nt}) \in \mathcal{X}$ the probability distribution used to draw J_t, conditionally to the past

$$\text{Regret}: R_T = \sum_{t=1}^T \ell_{J_t,t} - \min_k \sum_{t=1}^T \ell_{kt} = \left[\sum_{t=1}^T \ell_{J_t,t} - \sum_{t=1}^T \sum_{j=1}^N p_{jt}\ell_{jt}\right] + \left[\sum_{t=1}^T \sum_j p_{jt}\ell_{jt} - \min_k \sum_{t=1}^T \ell_{kt}\right]$$

This can be controlled independently of the probability distributions chosen

We already learned how to control this term! we denote it by $\overline{R}_T$ below

The information available at the begining of round t is $(\ell_s, p_s, J_s)_{s\leq t-1}$

We denote $\mathcal{F}_{t-1} = \sigma\{(\ell_s, p_s, J_s)_{s\leq t-1}\}$: ℓ_t and p_t are $\mathcal{F}_{t-1}$-measurable while J_t is drawn at random using an auxiliary randomization $U_t \sim \mathcal{U}_{[0,1]}$, independent from $\mathcal{F}_{t-1}$.

Then: $$\mathbb{E}[\ell_{J_t,t} \mid \mathcal{F}_{t-1}] = \sum_{j=1}^N p_{jt}\ell_{jt}$$ (J_t is not fixed by the conditioning, only its distribution p_t is so.)

↳ Expected regret (conditionally expected regret) $\overline{R}_T = \sum_{t,j} p_{jt}\ell_{jt} - \min_k \sum_{t=1}^T \ell_{kt}$

We already saw that we could ensure $\overline{R}_T \leq O\left((M-m)\sqrt{T \ln N}\right)$ if $\ell_{jt} \in [m, M]$ $\forall j, t$

↳ Martingale $$S_T = \sum_{t=1}^T \ell_{J_t,t} - \sum_{t=1}^T \sum_j p_{jt}\ell_{jt}$$

The Hoeffding-Azuma inequality ensures that if $\ell_{jt} \in [m, M]$ $\forall j \forall t$, then, no matter which p_t were selected

with $X_t = \ell_{J_t,t}$, $b_t = M$ and $a_t = m$, $\mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = \sum_j p_{jt}\ell_{jt}$

$$\mathbb{P}\left\{S_T \leq (M-m)\sqrt{\frac{T}{2}\ln 1/\delta}\right\} \geq 1-\delta$$

Conclusion: $\forall T$, with probability at least $1-\delta$, $$R_T \leq \overline{R}_T + (M-m)\sqrt{\frac{T}{2}\ln\frac{1}{\delta}}$$

E.g. with the fully adaptive version of EWA:

$\forall T$, $\forall \delta \in (0,1)$, with probability at least $1-\delta$, $R_T \leqslant (M-m)\sqrt{T}\left(\sqrt{\ln N} + \sqrt{\frac{1}{2}\ln\frac{1}{\delta}}\right) + (M-m)\left(2 + \frac{4}{3}\ln N\right)$

This is called a high probability bound; it is non-asymptotic → **Exercise:** Can you get a high probability bound of the form: $\forall \delta \in (0,1)$, with proba $\geqslant 1-\delta$, $\forall T$, $R_T \leqslant \ldots$?

Consequence: Asymptotic almost-sure bound.

The Borel-Cantelli lemma, using $\delta_T = 1/T^2$, ensures that

$$\mathbb{P}\left(\limsup_{\text{limsup of events}} \left\{ R_T > (M-m)\sqrt{T}\left(\sqrt{\ln N} + \sqrt{\ln T}\right) + (M-m)\left(2 + \frac{4}{3}\ln N\right) \right\}\right) = 0$$

We denote $\rho(T)$ this quantity: $\rho(T) \sim (M-m)\sqrt{T \ln T}$

That is, almost-surely

$R_T / \rho(T) > 1$ for finitely many T

thus, $\displaystyle\limsup_{T\to+\infty} \frac{R_T}{\rho(T)} \leqslant 1$ a.s. or equivalently,

limsup of a sequence of numbers

$$\limsup_{T\to+\infty} \frac{R_T}{(M-m)\sqrt{T\ln T}} \leqslant 1 \quad \text{a.s.}$$

Exercise: [To be stated in a more detailed way on the next page]

Show that we actually have $\displaystyle\limsup_{T\to+\infty} \frac{R_T}{(M-m)\sqrt{T\ln(\ln T)}} \leqslant C$ a.s. where C is a constant

(a rate which should be reminiscient of the law of the iterated logarithm.)

and I should have started with that...

Note: Of course, since $\mathbb{E}[S_T]=0$, we have $\mathbb{E}[R_T] = \mathbb{E}[\overline{R}_T]$

Because we have deterministic bounds on $\overline{R}_T$, we get bounds on $\mathbb{E}[R_T]$. But this doesn't tell us much on R_T, this is why we prefer our high-probability bounds.

 Lecture 2 — Winter 2022 term Lecture 2 — Winter 2022 term

Exercise [Full Statement]

(1) Remind yourself of Doob's martingale inequality (actually: inequalities – there are two of them, but we'll need only the most famous one).

(2) Show the following MAXIMAL version of the Hoeffding-Azuma inequality:

$\forall \delta \in (0,1)$, with probability at least $1-\delta$,

$$\max_{t \leq T} \left\{ \sum_{s=1}^{t} X_s - \sum_{s=1}^{t} \mathbb{E}[X_s \mid \mathcal{F}_{s-1}] \right\} \leq \sqrt{\frac{\sum_{t=1}^{T} (b_t - a_t)^2}{2} \ln \frac{1}{\delta}}$$

(3) Show that for any algorithm with expected regret $\overline{R}_T$ less than something of order $(M-m)\sqrt{T \ln N}$, the corresponding randomized algorithm has a regret R_T such that

For all strategies of the opponent picking losses $\ell_{j,t} \in [m, M]$,

$$\limsup_{T \to +\infty} \frac{R_T}{(M-m)\sqrt{T \ln(\ln T)}} \leq C \quad \text{a.s.}$$

where C is a universal constant (propose a numerical value).

(4) Is this C optimal? (Consider the law of the iterated logarithm as a basis for your discussion.)

Hint for (3): Consider the regimes $\{2^r+1, \ldots, 2^{r+1}\}$ for $r = 1, 2, \ldots$ and pick $\delta_r = 1/r^2$ for the application of the Borel-Cantelli lemma. (Cf. doubling trick!)