Sequential Learning, Sequential Optimization (= online machine learning)

How to reach me:

Gilles Stoltz : gilles.stoltz@math.u-psud.fr

Lecture notes posted on my webpage (Google search with "Gilles Stoltz")

↓

Go there NOW & register

Tentative schedule: Wednesdays at 2pm

January 20, 27
February 3, 10, 17
Vacation week: February 24
March 3, 10

Homework assignment #1

And the rest to be decided in due time:
- Class on March 17?
- Final exam on March 24 or 31 or none?
- Second homework?

Methodology: For each (sub)topic

1. Definition of the setting and of an objective (eg: regret)
2. Statement + analysis of an algorithm, preferably computationally efficient (eg: regret upper bound)
3. Lower bound on the objective, holding for any algorithm

Reminder: no simulation, no application to real data, 100% theoretical course ("definition – theorem – proof" style)

& no statistics background needed, probability students VERY welcome

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_j: 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:** $p_1 = \left(\frac{1}{N}, \dots \frac{1}{N}\right)$

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

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

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

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

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

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

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

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 $q \in \mathcal{X}$, 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, $$\sum_j p_{jt} \ell_{jt} \leq -\frac{1}{\eta} \underbrace{\ln \sum_j p_{jt} e^{-\eta \ell_{jt}}}_{= \ln \frac{\sum_j \exp(-\eta \sum_{s=1}^{t} \ell_{js})}{\sum_k \exp(-\eta \sum_{s=1}^{t-1} \ell_{ks})}} + \frac{\eta}{8}(M-m)^2$$

↑ by definition of the algorithm

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

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

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

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

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 $\eta\in\mathbb{R}$

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

(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

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

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

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

* 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 $\eta^* = \frac{1}{(M-m)}\sqrt{\frac{8\ln N}{T}}$ ← 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:

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

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 $\eta > 0$

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

given your choice for η

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

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

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 $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$)

$\leq O(1/\sqrt{T}) + \sup_{y\in\mathcal{Y}} \min_k f(x^{(k)}, y)$

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 $\mathcal{X}$:

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

is clearly $\geq$? $\sup_y \inf_x f(x,y)$ but maybe $\leq \sup_y \inf_x f(x,y)$ 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 $d_{\mathcal{Y}}$
- $f : \mathcal{X}\times\mathcal{Y} \to [0,M]$ uniformly continuous :

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

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

- 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 $\delta>0$:

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

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...)

Strongly convex loss functions → Same predictor, faster rates

(Still Setting #2, but ℓ_t only assumed to be convex, not linear)

Exp-concavity : The chosen loss functions ℓ_t are η-exp-concave if

$\forall t,$ $e^{-\eta \ell_t}$ is concave on $\mathcal{X}$

Remarks :

- $x \mapsto x^\alpha$ is concave and increasing for $0 < \alpha \leq 1$; thus η-exp-concave functions are also η'-exp-concave for all $0 < \eta' \leq \eta$
- $-\ln$ is convex and decreasing : exp-concave functions are in particular convex functions

Examples :

- Investment in the stock market, without transaction costs

1-exp-concave { $\ell_t(p) = -\ln\left(\sum_{j=1}^N p_j x_{jt}\right)$

where x_{jt} denotes the multiplicative evolution of stock j

The $-\ln$ comes because we consider a cumulative loss (not a reward)

- The square loss $\ell_t(p) = \left(\sum_j p_j f_{jt} - y_t\right)^2$

If $f_{jt}, y_t \in [0,B]$ $\forall j, \forall t$, then the ℓ_t are $1/2B^2$-exp-concave; it suffices to prove that $\forall y \in [0,B]$, $\psi_y : x \in [0,B] \mapsto e^{-\eta (x-y)^2}$ is concave for $\eta = 1/2B^2$

But $\psi_y'(x) = -2\eta(x-y)\, e^{-\eta(x-y)^2}$

$\psi_y''(x) = \left(4\eta^2(x-y)^2 - 2\eta\right) e^{-\eta(x-y)^2}$ is ≤ 0 as soon as $2\eta(x-y)^2 \leq 1$, which is guaranteed by $\eta \leq 1/2B^2$.

Same predictor : with $\ell_{jt} \overset{\text{def.}}{=} \ell_t(f_j)$; ie $p_{jt} \propto e^{-\eta \sum_{s=1}^{t-1} \ell_{js}}$

Theorem: If $\forall t$, ℓ_t is η-exp-concave, then the exponentially weighted average predictor used with this η is such that

(for all strategies to pick the ℓ_t)

$$\sum_{t=1}^{T} \ell_t(p_t) - \min_{k=1,\dots,N} \sum_{t=1}^{T} \ell_t(\delta_k) \leq \frac{\ln N}{\eta}$$

Ex: For all sequences of bounded forecasts and observations $\forall j,t \quad f_{jt} \in [0,B]$ and $\forall t, \ y_t \in [0,B]$,

$$\frac{1}{T} \sum_{t=1}^{T} \left(\sum_j p_{jt} f_{jt} - y_t \right)^2 \leq \frac{2B^2 \ln N}{T} + \min_{k=1,\dots,N} \frac{1}{T} \sum_{t=1}^{T} \left(f_{kt} - y_t \right)^2$$

In particular, taking $\sqrt{\ }$ and using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$:

$\forall$ sequences, RMSE of meta-predictor on stages $1,\dots,T$ $\leq$ RMSE of best predictor $+ B\sqrt{\frac{2 \ln N}{T}}$.

↳ Hence the naive "prediction of individual sequences" or "robust aggregation of forecasts"

Proof:

$$\ell_t(p_t) = -\frac{1}{\eta} \ln \underbrace{e^{-\eta \ell_t(p_t)}}_{\geq \sum_j p_{jt} e^{-\eta \ell_{jt}}} \leq -\frac{1}{\eta} \ln \sum_j p_{jt} e^{-\eta \ell_{jt}}$$

Same telescoping argument:

$$\sum_{t=1}^{T} \ell_t(p_t) \leq -\frac{1}{\eta} \ln \frac{\sum_{j=1}^{N} e^{-\eta \sum_{t=1}^{T} \ell_{jt}}}{N}$$

and same way of concluding as before. ∎

What about the more challenging comparison to convex combinations ?

$$\inf_{p \in X} \sum_{t=1}^{T} \ell_t(p)$$

Exp-concavity & comparison to convex losses

Algorithm:

(continuous EWA prediction)

$$p_t = \frac{\int_\chi p\, e^{-\eta \sum_{s=1}^{t-1} \ell_s(p)}\, d\mu(p)}{\int_\chi e^{-\eta \sum_{s=1}^{t-1} \ell_s(p)}\, d\mu(p)} \overset{\text{not.}}{=} \int_\chi p\, d\mu_t(p)$$

where μ is the uniform (Lebesgue) measure on χ.

implicitly defining a measure μ_t here

Remarks:

- $p_1 = (1/N \dots 1/N)$
- how to better grasp μ? χ is included in an affine hyperspace, in which it has a positive Lebesgue measure → μ is a conditional measure of that
- one can show that if $X_1 \dots X_{N-1}$ iid $\sim \mathcal{U}_{[0,1]}$ then $0 \leq X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(N-1)} \leq 1$ (order statistics) define N segments
$$\left(X_{(1)},\ X_{(2)} - X_{(1)},\ \dots,\ 1 - X_{(N-1)}\right)$$
whose joint law is μ.
- how to compute p_t?
  - ↳ grid arguments (with δ intervals) have a complexity of $O(\delta^{-N+1})$
  - ↳ Monte-Carlo methods? Simulate $P_1 \dots P_m$ iid $\sim \mu$ and use $\int_\chi g(p)\, d\mu(p) \simeq \frac{1}{m} \sum_{j=1}^{m} g(P_j)$
  - ↳ MCMC methods, see Kompala 1996.

in any case, a nightmare in practice (I had to implement it once, see Stoltz & Lugosi 2005)

Theorem: For all sequences ℓ_t of η-exp-concave functions, the continuous EWA predictor satisfies:

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

PROOF:

$$\ell_t(p_t) = -\frac{1}{\eta} \ln e^{-\eta \ell_t(p_t)}$$

$$= -\frac{1}{\eta} \ln e^{-\eta \ell_t\left(\int_{\mathcal{X}} p \, d\mu_t(p)\right)}$$

$$\geq \int_{\mathcal{X}} e^{-\eta \ell_t(p)} d\mu_t(p) \quad \text{by exp-concavity}$$

thus

$$\ell_t(p_t) \leq -\frac{1}{\eta} \ln \int_{\mathcal{X}} e^{-\eta \ell_t(p)} d\mu_t(p)$$

$$= -\frac{1}{\eta} \ln \frac{\int_{\mathcal{X}} e^{-\eta \sum_{s=1}^{t} \ell_s(p)} d\mu(p)}{\int_{\mathcal{X}} e^{-\eta \sum_{s=1}^{t-1} \ell_s(p)} d\mu(p)}$$

Telescoping argument:

$$\sum_{t=1}^{T} \ell_t(p_t) \leq \cancel{\frac{\ln 1}{\eta}} - \frac{1}{\eta} \ln \int_{\mathcal{X}} e^{-\eta \sum_{t=1}^{T} \ell_t(p)} d\mu(p)$$

[Fix $\delta > 0$, let p_δ^* be s.t. $\inf_p \sum_{t=1}^{T} \ell_t(p) \leq \delta + \sum_{t=1}^{T} \ell_t(p_\delta^*)$

‖ why is the inf not a min? ℓ_t is convex thus continuous on the interior of $\mathcal{X}$ but can be discontinuous on its border (1D-example: $x \in [0,1] \mapsto x^2 + \mathbb{1}_{x=0}$)

$$\Delta_{\delta,\varepsilon}^* = \left\{ (1-\varepsilon) p_\delta^* + \varepsilon r, \; r \in \mathcal{X} \right\} \quad \text{for } \varepsilon \in]0,1[$$

→ for such $p = (1-\varepsilon) p_\delta^* + \varepsilon r$, we have by exp-concavity:

$\forall t$, $e^{-\eta \ell_t(p)} \geq (1-\varepsilon) e^{-\eta \ell_t(p_\delta^*)} + \varepsilon \underbrace{e^{-\eta \ell_t(r)}}_{\geq 0}$

→ also, $\mu(\Delta_{\delta,\varepsilon}^*) = \varepsilon^{N-1}$ (and ambient dimension $N-1$)

cf. properties of the Lebesgue measure

Thus

$$\int_{\mathcal{X}} e^{-\eta \sum_{t=1}^{T} \ell_t(p)} \, d\mu(p) \geqslant (1-\varepsilon)^T \, \varepsilon^{N-1} \, e^{-\eta \sum_{t=1}^{T} \ell_t(p_\delta^*)}$$

Substituting in the bound:

$$\sum_{t=1}^{T} \ell_t(p_t) \leqslant -\frac{1}{\eta} \ln\left((1-\varepsilon)^T \varepsilon^{N-1}\right) + \underbrace{\sum_{t=1}^{T} \ell_t(p_\delta^*)}_{\leqslant \delta + \inf_{p \in \mathcal{X}} \sum_{t=1}^{T} \ell_t(p)}$$

Let $\delta \downarrow 0$, we have proved:

$$\sum_{t=1}^{T} \ell_t(p_t) - \inf_{p \in \mathcal{X}} \sum_{t=1}^{T} \ell_t(p) \leqslant \inf_{\varepsilon \in]0,1[} -\frac{1}{\eta} \ln\left((1-\varepsilon)^T \varepsilon^{N-1}\right)$$

Choosing, e.g., $\varepsilon = 1/(T+1)$:

$$1/(1-\varepsilon)^T = \left(1 - 1/(T+1)\right)^{-T} = \left(T/(T+1)\right)^{-T} = \left(1 + 1/T\right)^T$$

$$= \exp\Big(T \underbrace{\ln\left(1 + 1/T\right)}_{\leqslant 1/T}\Big) \leqslant e$$

hence the stated bound. ↵

A better bound?

$$B(\varepsilon) = T \ln(1-\varepsilon) + (N-1) \ln \varepsilon \qquad B'(\varepsilon) = \frac{-T}{1-\varepsilon} + \frac{N-1}{\varepsilon}$$

$B''(\varepsilon) < 0$ thus the ε^* s.t. $B'(\varepsilon^*) = 0$ is the global minimum of B

We have $B'(\varepsilon^*) = 0 \iff \frac{\varepsilon^*}{N-1} = \frac{1}{T} - \frac{\varepsilon^*}{T} \iff \varepsilon^* = \frac{1/T}{1/T + 1/(N-1)} = \frac{N-1}{T+N-1}$

We thus get the sharper final bound

$$\frac{1}{\eta}\left(T \ln\left(\frac{T+N-1}{T}\right) + (N-1) \ln\left(\frac{T+N-1}{N-1}\right)\right) = \frac{T+N-1}{\eta} H\left(\frac{N-1}{T+N-1}\right)$$

$$= \frac{1}{\eta}\Big((N-1) \underbrace{\ln\left(1 + \frac{N-1}{T}\right)}_{\leqslant (N-1)/T} + (N-1) \ln\left(1 + \frac{T}{N-1}\right)\Big)$$

where $H(x) = x \ln x + (1-x) \ln(1-x)$ is the binary entropy

* However * both bounds are $\sim \frac{N-1}{\eta} \ln T$ as $T \to +\infty$.

↳ And the second bound seems less readable to me...

Third case : Convex Loss Functions

$\ell_t : X \to [m,M]$ convex

1. Comparison to the best individual expert

Regret $R_T = \sum_{t=1}^T \ell_t(p_t) - \min_{i=1,\dots,N} \sum_{t=1}^T \ell_t(\delta_i)$ where δ_i Dirac mass at i

is bounded by convexity by

$$\tilde{R}_T \leq \underbrace{\sum_{t=1}^T \sum_{j=1}^N p_{jt}\, \ell_{jt}}_{\text{linear upper bound, which we know how to control with EWA}} - \min_{i=1,\dots,N} \sum_{t=1}^N \ell_{it}$$

where we denoted $\ell_{jt} = \ell_t(\delta_j)$

2. Comparison to the best convex vector

Exercise #3 : An inefficient algorithm to do so : the continuous EWA predictor

(↳ Later we'll see a more efficient way (algorithm with a better computational complexity) + a better regret bound)

1) Show that the continuous EWA strategy is such that :

Against all sequences $\ell_t : X \to [m,M]$ of convex functions,

$$\sum_{t=1}^T \ell_t(p_t) - \inf_{q \in X} \sum_{t=1}^T \ell_t(q) \leq \frac{\eta}{8}(M-m)^2 T + \inf_{\varepsilon \in \Delta(\Omega)} \frac{1}{\eta} \ln\left(\varepsilon^{N-1} e^{-\eta \varepsilon T (M-m)}\right)$$

2) When T, M and m are known, which η is a good choice to minimize the obtained upper bound, and can you provide an easy-to-read final upper bound ?