Sequential optimization for general convex sets.

We so far focused our attention on the simplex

$$\mathcal{X} = \left\{ (p_1 \dots p_N) : \forall j,\ p_j \geq 0 \text{ and } \sum_k p_k = 1 \right\} \subset \mathbb{R}^N$$

Let's now deal with more general convex sets $\mathcal{C} \subset \mathbb{R}^N$, which we assume to be CLOSED.

Protocol: At each round $t = 1, 2, \dots$

1. The statistician picks $x_t \in \mathcal{C}$ while the opponent picks simultaneously a convex and DIFFERENTIABLE function $\ell_t : \mathcal{C} \to \mathbb{R}$
2. x_t and ℓ_t are revealed

Aim: Control $R_T = \sum_{t=1}^T \ell_t(x_t) - \inf_{x \in \mathcal{C}} \sum_{t=1}^T \ell_t(x)$ (well, the inf is a min here)

Algorithm: Online gradient descent with fixed learning rate $\eta > 0$ (called $\mathrm{OGD}(\eta)$)

- Play any $x_1 \in \mathcal{C}$
- For $t = 2, 3, \dots$,

$$x_t = \Pi_{\mathcal{C}} \left(x_{t-1} - \eta \nabla \ell_{t-1}(x_{t-1}) \right)$$

where $\Pi_{\mathcal{C}}$ is the Euclidean projection onto $\mathcal{C}$ (well defined because $\mathcal{C}$ is closed).

$\|\cdot\|$ is the Euclidean norm.

Theorem: Assume that $\max_{t \leq T} \sup_{x \in \mathcal{C}} \|\nabla \ell_t(x)\| \leq G$ [bounded gradients]

and that $\mathcal{C}$ is compact, i.e., $\sup_{x, x' \in \mathcal{C}} \|x - x'\| \leq D$ [bounded diameter]

Then $\mathrm{OGD}(\eta)$ is such that $\sum_{t=1}^T \ell_t(x_t) - \min_{x \in \mathcal{C}} \sum_{t=1}^T \ell_t(x) \leq \frac{D^2}{2\eta} + \frac{\eta}{2} G^2 T$

(inf achieved by continuity + compactness)

In particular, for $\eta = \frac{D}{G\sqrt{T}}$, the bound equals $DG\sqrt{T}$.

Remark: In the case with EWA, we were considering a linear function:

$$\forall p \in \mathcal{X}, \quad \ell_t(p) = \sum_{j=1}^{N} p_j \ell_{jt}$$

↳ OGD would also be applicable in this setting, though it leads to a suboptimal bound:

diameter of $\mathcal{X}$: $D^2 = 2$

$$\forall p,q \in \mathcal{X}, \quad \sum_j (p_j - q_j)^2 \leq \sum_j |p_j - q_j| \leq \sum_j (p_j + q_j) = 2$$

equality achieved if p, q are Dirac masses at $i \neq k$

bound on the gradients: if, eg, $\ell_{jt} \in [-M, M] \ \forall j,t$,

then $\nabla \ell_t(p) = \begin{pmatrix} \ell_{1t} \\ \ell_{Nt} \end{pmatrix}$

and $\|\nabla \ell_t(p)\|^2 \leq N M^2 = G^2$

We get the bounds:

- for OGD: $DG\sqrt{T} = M\sqrt{2NT}$
- for EWA: $(M-(-M))^2 \sqrt{\frac{T}{2} \ln N} = M\sqrt{8T \ln N}$

The dependency in N is suboptimal for OGD.

Proof (of the theorem): By convexity (« inégalité des pentes » in French):

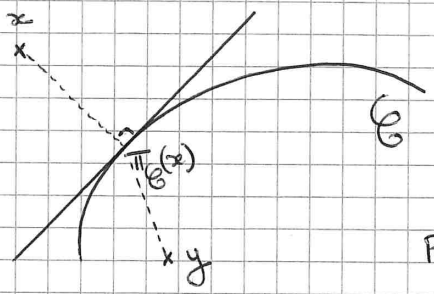
$$\forall x \in \mathcal{C}, \forall t, \quad \ell_t(x_t) - \ell_t(x) \leq \nabla \ell_t(x_t) \cdot (x_t - x)$$

Summing over t,

$$R_T = \sup_{x \in \mathcal{C}} \left\{ \sum_{t=1}^{T} \ell_t(x_t) - \sum_{t=1}^{T} \ell_t(x) \right\}$$

$$\leq \sup_{x \in \mathcal{C}} \left\{ \sum_{t=1}^{T} \nabla \ell_t(x_t) \cdot (x_t - x) \right\}$$

Now, we will need some facts about projections on CLOSED convex sets:



Def: $\Pi_{\mathcal{C}}(x)$ is the (unique) $y^* \in \mathcal{C}$ with $\|x - y^*\| \leq \|x - y\|$ $\forall y \in \mathcal{C}$

Property 1: if $x \notin \mathcal{C}$, then $\forall y \in \mathcal{C}$, $(y - \Pi_{\mathcal{C}}(x)) \cdot (x - \Pi_{\mathcal{C}}(x)) \leq 0$

Property 2: $\forall y \in \mathcal{C}$, $\|y - \Pi_{\mathcal{C}}(x)\| \leq \|y - x\|$

Proof: (Property 1): $t \in [0,1]$, $y_t = \Pi_{\mathcal{C}}(x) + t(y - \Pi_{\mathcal{C}}(x)) \in \mathcal{C}$ by convexity

$$\|x - y_t\|^2 = \|x - \Pi_{\mathcal{C}}(x)\|^2 + t^2 \|y - \Pi_{\mathcal{C}}(x)\|^2 - 2t (y - \Pi_{\mathcal{C}}(x)) \cdot (x - \Pi_{\mathcal{C}}(x)) \geq \|x - \Pi_{\mathcal{C}}(x)\|^2$$

simplify the $\|x - \Pi_{\mathcal{C}}(x)\|^2$, divide by t and $t \to 0$

(Property 2) $$\|y - x\|^2 = \|y - \Pi_{\mathcal{C}}(x) + \Pi_{\mathcal{C}}(x) - x\|^2 = \|y - \Pi_{\mathcal{C}}(x)\|^2 + \|x - \Pi_{\mathcal{C}}(x)\|^2 + 2(y - \Pi_{\mathcal{C}}(x)) \cdot (\Pi_{\mathcal{C}}(x) - x)$$

both are ≥ 0 (by Prop 1)

In particular, for $t \geq 1$:

$$\forall x \in \mathcal{C}, \quad \|x_{t+1} - x\|^2 \overset{\text{def of OGD}}{=} \|\Pi_{\mathcal{C}}(x_t - \eta \nabla \ell_t(x_t)) - x\|^2$$

$$\overset{\text{Property 2}}{\leq} \|x_t - \eta \nabla \ell_t(x_t) - x\|^2$$

$$= \|x_t - x\|^2 + \eta^2 \|\nabla \ell_t(x_t)\|^2 - 2\eta \nabla \ell_t(x_t) \cdot (x_t - x)$$

That is, $\forall t, \forall x \in \mathcal{C}$,

$$\nabla \ell_t(x_t) \cdot (x_t - x) \leq \frac{1}{2\eta} \left(\|x_t - x\|^2 - \|x_{t+1} - x\|^2 \right) + \frac{\eta}{2} \underbrace{\|\nabla \ell_t(x_t)\|^2}_{\leq G^2 \text{ by the assumption of bounded gradients}}$$

Summing over t:

$$\sum_{t=1}^{T} \nabla \ell_t(x_t) \cdot (x_t - x) \leq \frac{1}{2\eta} \left(\underbrace{\|x_1 - x\|^2}_{\leq D^2} - \underbrace{\|x_{T+1} - x\|^2}_{\geq 0} \right) + \frac{\eta}{2} T G^2$$

Adaptation to time T (and other parameters) / for OGD and EWA

The algorithms discussed so far:

EWA with fixed learning rate $\eta > 0$
OGD with fixed step size $\eta > 0$

required the knowledge of T and of

the range $[m, M]$ of the losses for EWA
the bound G on the gradients of the losses for OGD

We now study adaptation to these parameters.

There exists a general trick, called the doubling trick, which we will detail in 4 pages from now for EWA.

Another popular way of being adaptive is to tune η according to the past observations:

Adaptive OGD:

- Play an arbitrary $x_1 \in \mathcal{C}$, pick $\eta_1 > 0$
- For $t \geq 2$, play $x_t = \Pi_{\mathcal{C}}\left(x_{t-1} - \eta_{t-1} \nabla \ell_{t-1}(x_{t-1})\right)$
  pick $\eta_t > 0$

Theorem: If Adaptive OGD picks a non-increasing sequence $(\eta_t)_t$ with $\eta_t > 0$ $\forall t$,

then: $$R_T \leq \frac{D^2}{2\eta_T} + \frac{1}{2} \sum_{t=1}^{T} \eta_t \|\nabla \ell_t(x_t)\|^2$$

with the same notation as in the theorem for OGD(η)

Corollary: Picking $$\eta_t = \begin{cases} 1 & \text{if } V_t = \sum_{s=1}^{t} \|\nabla \ell_s(x_s)\|^2 < D^2 \\ D/\sqrt{V_t} & \text{if } V_t \geq D^2 \end{cases}$$

Adaptive OGD achieves the bound $$\max\left\{ D^2, \frac{3}{2} D \sqrt{\sum_{t=1}^{T} \|\nabla \ell_t(x_t)\|^2} \right\} = O\left(DG\sqrt{T}\right)$$

without any prior knowledge of neither T nor G

(of course, $\mathcal{C}$ thus D are known).

Proof of the theorem and of the corollary:

Th: We extract from the proof of $OGD(\eta)$ that

$$\forall x \in \mathcal{C}, \quad \nabla \ell_t(x_t) \cdot (x_t - x) \leq \frac{1}{2\eta_t}\left(\|x_t - x\|^2 - \|x_{t+1} - x\|^2 \right) + \frac{\eta_t}{2} \|\nabla \ell_t(x_t)\|^2$$

Summing over t:

$$\forall x \in \mathcal{C}, \quad \sum_{t=1}^{T} \ell_t(x_t) - \sum_{t=1}^{T} \ell_t(x) \leq \sum_{t=1}^{T} \nabla \ell_t(x_t) \cdot (x_t - x)$$

$$\leq \sum_{t=1}^{T} \frac{1}{2\eta_t}\left(\|x_t - x\|^2 - \|x_{t+1} - x\|^2 \right) + \frac{1}{2} \sum_{t=1}^{T} \eta_t \|\nabla \ell_t(x_t)\|^2$$

We need only to show that this sum is $\leq D^2/2\eta_T$

Abel's transform:

$$\sum_{t=1}^{T} \frac{1}{2\eta_t}\left(\|x_t - x\|^2 - \|x_{t+1} - x\|^2 \right)$$

$$= \frac{1}{2\eta_1} \underbrace{\|x_1 - x\|^2}_{\leq D^2} + \sum_{t=2}^{T} \underbrace{\left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right)}_{\geq 0 \text{ because } (\eta_t) \text{ is non-increasing}} \underbrace{\|x_t - x\|^2}_{\leq D^2} - \underbrace{\frac{1}{2\eta_T} \|x_{T+1} - x\|^2}_{\leq 0}$$

$$\leq \frac{D^2}{2\eta_T} \quad \text{after telescoping}$$

Cor: Recall that we picked:

$$\forall t \geq 1, \quad \eta_t = \begin{cases} 1 & \text{if } V_t = \sum_{s=1}^{t} \|\nabla \ell_s(x_s)\|^2 < D^2 \\ D/\sqrt{V_t} & \text{if } V_t \geq D^2 \end{cases}$$

Pay attention to the indexation: η_t is used to pick x_{t+1} and may thus depend on the past till and including round t

(Some practice actually explains why this is a natural choice → we replace the theoretically optimal but

impossible in practice choice $\eta_t \equiv D/V_T$ by its "sequential estimation"; and use a threshold to avoid too large values of η_t for small t.)

The regret bound reads $\frac{D^2}{2\eta_T} + \frac{1}{2}\sum_{\substack{t: \\ V_t < D^2}} \|\nabla \ell_t(x_t)\|^2 + \frac{1}{2}\sum_{\substack{t: \\ V_t \geq D^2}} \frac{D}{\sqrt{V_t}} \|\nabla \ell_t(x_t)\|^2$

We denote τ the largest $t \leq T$ such that $V_t < D^2$. Then, if $\tau < T$, we have $V_{\tau+1} \geq D^2$ (possibly $t=0$ if $V_1 \geq D^2$)

The above bound can be rewritten as:

$$\frac{D^2}{2\eta_T} + \underbrace{\frac{1}{2}\sum_{t: V_t < D^2} \|\nabla \ell_t(x_t)\|^2}_{= \frac{V_\tau}{2} < D^2/2} + \underbrace{\frac{1}{2}\sum_{t=\tau+1}^{T} \frac{D}{\sqrt{V_t}} \|\nabla \ell_t(x_t)\|^2}_{\text{possibly void sum (if } \tau = T)}$$

$$= \frac{1}{2}\sum_{t=\tau+1}^{T} \frac{D}{\sqrt{V_t}} (V_t - V_{t-1}) = \frac{1}{2}\sum_{t=\tau+1}^{T} \frac{D}{\sqrt{V_t}} \underbrace{\left(\sqrt{V_t} + \sqrt{V_{t-1}}\right)}_{\leq 2\sqrt{V_t}} \left(\sqrt{V_t} - \sqrt{V_{t-1}}\right)$$

$$\leq D \sum_{t=\tau+1}^{T} \left(\sqrt{V_t} - \sqrt{V_{t-1}}\right)$$

$$= D\left(\sqrt{V_T} - \sqrt{V_\tau}\right)$$

We conclude by distinguishing two cases:

* $\tau = T$: $\eta_T = 1$ and the bound is $\leq \frac{D^2}{2} + \frac{D^2}{2} = D^2$

* $\tau < T$: $\eta_T = D/\sqrt{V_T}$ and the bound is

$$\frac{D^2}{2D/\sqrt{V_T}} + \underbrace{\frac{V_\tau}{2}}_{\leq \frac{D\sqrt{V_\tau}}{2} \text{ as } V_\tau < D^2} + D\sqrt{V_T} - D\sqrt{V_\tau} \leq \frac{3}{2} D\sqrt{V_T}$$

↳ A readable and simple final bound is thus the one we indicated:

$$\max\left\{D^2, \frac{3}{2}D\sqrt{V_T}\right\} = \max\left\{D^2, \frac{3}{2}D\sqrt{\sum_{t=1}^{T} \|\nabla \ell_t(x_t)\|^2}\right\}$$

Calibration of η for the EWA strategy.

Reminder: by considering subgradients and pseudo-losses, we can reduce the case of convex losses to the one of linear losses. It thus suffices to explain how to sequentially calibrate η in the latter case.

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

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

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

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

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

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

Calibration of η for the EWA strategy: much harder!

The cheap way: the doubling trick.

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

RESTART EWA with $\eta_r = \frac{1}{M-m}\sqrt{\frac{8 \ln N}{2^r}}$ for rounds $t = 2^r+1, \dots, 2^{r+1}$ (2^r such rounds)

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

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

We have

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

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

(2(M-m): rounds $t=1$ and $t=2$; max of a sum $\leq$ sum of the max.)

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 $\leq \frac{\ln N}{\eta_{r_T}} + \eta_{r_T}(M-m)^2 \frac{(T-2^{r_T})}{8}$
$\leq \frac{\ln N}{\eta_{r_T}} + \eta_{r_T}(M-m)^2 \frac{2^{r_T}}{8}$ ($T \leq 2^{r_T+1}$)
$= (M-m)\sqrt{\frac{2^{r_T}}{2}\ln N}$

We have proved:

$$R_T \leq 2(M-m) + \left(\sum_{r=1}^{r_T}\sqrt{2^r}\right)\sqrt{\frac{\ln N}{2}} \times (M-m)$$

$$\sum_{r=1}^{r_T}\sqrt{2^r} = \sqrt{2} \times \frac{(\sqrt{2})^{r_T}+1}{\sqrt{2}-1}$$

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

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

the main difference w.r.t. bound when T is known is an additional $\sqrt{2}/\sqrt{2}-1 \approx 3{,}42$ which measures the price for the adaptivity in T if one uses the doubling trick

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 "estimate" 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, as we did for OGD :

The smarter way : past-data adaptive choice of η

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 : $p_{jt}(\eta) = \exp\left(-\eta \sum_{s=1}^{t} \ell_{js}\right) \Big/ \sum_{i=1}^{N} \exp\left(-\eta \sum_{s=1}^{t} \ell_{is}\right)$.

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 ....)

Not the interaction protocol is the same !!!

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 $\delta_t = \sum_{j=1}^{N} p_{jt} \ell_{jt} + \frac{1}{\eta_t} \ln\left(\sum_{j=1}^{N} p_{jt} e^{-\eta_t \ell_{jt}}\right)$

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

Exercise 1: (easy: see it as a warm-up!) 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?).

Proof (of the lemma):

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

The question is thus to control $\sum_{t=1}^{T} -\frac{1}{\eta_t} \ln\left(\sum_j p_{jt} e^{-\eta_t \ell_{jt}}\right)$ ↖ $\leq \frac{\ln N}{\eta} + \min_k \sum_{t=1}^{T} \ell_{kt}$?

↳ 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:

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

$$\geq_{\text{Jensen}} \left(\frac{1}{N}\sum_j p_{jt}^{\eta_{t+1}/\eta_t} e^{-\eta_{t+1}\ell_{jt}}\right)^{\eta_t/\eta_{t+1}}$$

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

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

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

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

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

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

$$= \frac{\ln N}{\eta_{T+1}}$$

$$= \frac{\ln N}{\eta_{T+1}} - \frac{1}{\eta_{T+1}} \ln\left(\sum_j e^{-\eta_{T+1} \sum_{s=1}^{T} \ell_{js}} \right)$$

$$\geq e^{-\eta_{T+1} \min_k \sum_{t=1}^{T} \ell_{kt}}$$

$$\leq \min_{k \leq N} \sum_{t=1}^{T} \ell_{kt}$$

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 $\eta'_{T+1} = \eta_T$)

Comments: 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).

Intuition: $e^x \simeq 1 + x + x^2/2$ when x is small

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

$\ln(1+u) \simeq u$ when u is small

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

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

Actually, we could do better:

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

$$\underset{\eta_t \text{ large}}{\approx} -\frac{1}{\eta_t}\ln\left(1 - \sum_j p_{jt}\left(\eta_t\left(\ell_{jt} - \sum_k p_{kt}\ell_{kt}\right)\right) + \frac{\eta_t^2}{2}\sum_j p_{jt}\left(\ell_{jt} - \sum_k p_{kt}\ell_{kt}\right)^2\right)$$

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

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

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

(In the case of costs, this argument was worked out in a slightly more involved way.)

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

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

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

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

↳ For a decade (2005 - 2014) this proof was somewhat wrong 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.

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

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

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

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

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

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

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

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

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

Note: $p_1 = (1/N, \dots 1/N)$ and $\eta_1 = +\infty$; since the value of η_1 is irrelevant this merely means that $\eta_1 \geq \eta_2$ 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 $\delta_1 > 0$.

Theorem: The above strategy ensures that:

$\forall\, m \leq M$ (not known by the algorithm), $\forall\, \ell_{jt} \in [m,M]$,

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

Proof: $\eta_t \searrow$ then $R_T \leq \frac{\ln N}{\eta_T} + \sum_{t=1}^T \delta_t$ ($\eta_1 = +\infty$ means that we do here $\eta_1 \to +\infty$)

By Bernstein's inequality (with $X_t = -\ell_{jt}$ with probability p_{jt} and $\eta_t > 0$): $\delta_t \leq \frac{1}{\eta_t} \frac{e^{\eta_t^2 (1-m)} - \eta_t (1-m) - 1}{(M-m)^2} v_t$ for $t \geq 2$

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

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

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

$= \frac{2}{\eta_t (M-m)} (M-m) \delta_t - (M-m) \delta_t \times f(\eta_t (M-m))$

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

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

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

Then, $\left(\sum_{t=1}^T \delta_t \right)^2 = \sum_{t=1}^T \left(\left(\sum_{s=1}^t \delta_s \right)^2 - \left(\sum_{s=1}^{t-1} \delta_s \right)^2 \right)$ (telescoping)

$= \sum_{t=1}^T \delta_t \left(\delta_t + 2 \sum_{s=1}^{t-1} \delta_s \right)$ ($a^2 - b^2 = (a+b)(a-b)$)

(δ_t: direct bound: this is $\leq M-m$; $\sum_{s=1}^{t-1} \delta_s$: by def of the algorithm, this is $\frac{\ln N}{\eta_t}$, even for $t=1$)

$\leq \sum_{t=1}^T \left((M-m)\delta_t + \frac{2\delta_t}{\eta_t} \ln N \right)$

$\leq v_t \ln N + \frac{2}{3}(M-m)\delta_t \ln N$

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

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

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

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

$$x \leq \frac{a + \sqrt{a^2 + 4b}}{2} \leq a + \sqrt{b} \qquad \text{using } \sqrt{t+t'} \leq \sqrt{t} + \sqrt{t'}. \ ]$$

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

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

Final bound: $$R_T \leq 2 \sum_{t=1}^{T} \delta_t.$$

Exercise 2. The proof above seems unnatural to me; I do not really see the link to the intuition I give you.

→ If you can find a more natural exposition (still starting from Bernstein's inequality), you will get bonus points at the exam!

Corollary: The algorithm above is such that

$\forall T$, $\forall m \leq M$ (not known by the algorithm), $\forall f_t \in [m, M]$,

$$\sum_{t=1}^{T} p_t f_t - \min_k \sum_{t=1}^{T} f_{kt} \leq (M-m)\sqrt{T \ln N} + (M-m)\left(2 + \frac{4}{3}\ln N\right)$$

↳ The algorithm is adaptive to T and to the bounds m and M.

Proof: V_t = some variance of a random variable $\in [m, M]$ $\leq \frac{(M-m)^2}{4}$, as in the proof of Hoeffding's lemma.

Exercise 3. 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 rate" results in classification/statistical

learning.

(1) Assume that losses ℓ_{jt} are non-negative: $\ell_{jt} \geq 0$, ie, $\ell_{jt} \in [0, M]$.
Show that $$\sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \left(\ell_{jt} - \sum_{k} p_{kt} \ell_{kt} \right)^2 \leq M \sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \ell_{jt}.$$

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

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

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

Exercise 4. « Impossible tuning »

Consider the PROD forecaster: $p_1 = (1/N, \dots 1/N)$
and for $t \geq 2$, $$p_{jt} = \prod_{s=1}^{t-1} (1 - \eta \ell_{js}) \Big/ \sum_{k=1}^{N} \prod_{s=1}^{t-1} (1 - \eta \ell_{ks})$$

It appears as a first-order approximation to EWA.

(1) Show that $$\sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \ell_{jt} - \sum_{t=1}^{T} \ell_{kt} \leq \frac{\ln N}{\eta} + \eta \sum_{t=1}^{T} \ell_{kt}^2$$
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:
$$\sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \ell_{jt} \leq \min_{k=1\dots N} \left\{ \sum_{t=1}^{T} \ell_{kt} + C \sqrt{\sum_{t=1}^{T} \ell_{kt}^2 \ln N} \right\}$$
(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 —

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!