Algorithm for a distribution-free bound.

We will study it in the following, more general, setting, called the adversarial setting:

For $t = 1, 2, \ldots$:

- The opponent picks $(g_{1t}, \ldots g_{Kt}) \in [0,1]^K$
  while, simultaneously,
- The decision-maker picks $I_t \in \{1..K\}$, possibly at random according to a distribution $p_t = (p_{1t} \ldots p_{Kt})$
  and thanks to an auxiliary randomization U_t
- The payoff of the decision-maker is $g_{I_t,t}$, and this is the only feedback she gets (she does not observe g_{jt} for $j \neq I_t$)

Aim: Control the regret $R_T = \max_{k=1\ldots K} \sum_{t=1}^{T} g_{kt} - \sum_{t=1}^{T} g_{I_t,t}$

- in high probability (preferred but difficult)
- in expectation: $\mathbb{E}[R_T]$

Reminder: each g_{kt} depends on $I_1 \ldots I_{t-1}$

Example: A stochastic opponent picks, once for all, $\nu_1 \ldots \nu_K \in \mathcal{P}([0,1])$ and draws $g_{kt} \sim \nu_k$ for all t, independently.

Then (as we already discussed):

$$\mathbb{E}[R_T] \geqslant T \max_{k=1..K} E(\nu_k) - \mathbb{E}\left[\sum_{t=1}^{T} g_{I_t,t}\right] = \bar{R}_T$$

Fundamental observations: Assume with no loss of generality that the opponent is deterministic.

Then g_{kt} is $\mathcal{F}_{t-1} = \sigma(I_1 .. I_{t-1}) = \sigma(U_1 .. U_{t-1})$ - measurable, and so is p_t.

Thus, in $\mathbb{E}[g_{I_t t} \mid \mathcal{F}_{t-1}]$, the only randomness to integrate over, is the random draw of I_t according to p_t thanks to U_t:

$$\mathbb{E}\left[g_{I_t,t} \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\sum_{k=1}^{K} g_{kt} \mathbb{1}_{\{I_t = k\}} \;\middle|\; \mathcal{F}_{t-1}\right]$$

$$= \sum_{k=1}^{K} g_{kt} \underbrace{\mathbb{E}\left[\mathbb{1}_{\{I_t = k\}} \mid \mathcal{F}_{t-1}\right]}_{= p_{kt} \text{ by the model}} = \sum_{k=1}^{K} p_{kt}\, g_{kt}$$

This leads to the following important observation: if $p_{kt} > 0$, then $\hat{\ell}_{kt} = \frac{(1-g_{I_t,t})}{p_{kt}} \mathbb{1}_{\{I_t = k\}}$ is a (conditionally) unbiased estimator of the pseudo-losses $\ell_{kt} = 1 - g_{kt}$

indeed, similar calculations show that

$$\mathbb{E}\left[\hat{\ell}_{kt} \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\frac{1-g_{kt}}{p_{kt}} \mathbb{1}_{\{I_t = k\}} \mid \mathcal{F}_{t-1}\right] = \frac{1-g_{kt}}{p_{kt}} \underbrace{\mathbb{E}\left[\mathbb{1}_{\{I_t=k\}} \mid \mathcal{F}_{t-1}\right]}_{= p_{kt}} \underset{\text{as } p_{kt} > 0}{=} 1 - g_{kt}$$

exponentially weighted averages

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

Set $p_1 = (1/K, \dots 1/K)$, draw $I_1 \sim p_1$, get $g_{I_1,1}$, compute the $\hat{\ell}_{k,1}$ for $k = 1 .. K$

For $t = 2, 3, \dots$

- set $p_{it} = \exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{is}\right) \Big/ \sum_{k=1}^{K} \exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{ks}\right)$

- draw $I_t \sim p_t$, get $g_{I_t,t}$, compute the $\hat{\ell}_{kt}$ for $k = 1, \dots K$

Theorem: For all opponents picking payoffs $g_{kt} \in [0,1]$, for all $T \geq 1$,

$$\overline{R}_T = \max_{k=1..K} \mathbb{E}\left[\sum_{t=1}^{T} g_{kt}\right] - \mathbb{E}\left[\sum_{t=1}^{T} g_{I_t,t}\right] \leq \frac{\ln K}{\eta} + \eta T \frac{K}{2}.$$

In particular, for $\eta = \sqrt{\frac{2K \ln K}{T}}$, we get $\overline{R}_T \leq \sqrt{2 TK \ln K}$

Issue 1: The horizon T is unknown in advance, and even worse, we may have $T \to +\infty$. We will need to study how to better tune η.

Issue 2: We don't deal yet with $\mathbb{E}[R_T]$ ... → we might do so in the exam!

Before we actually prove this theorem, let us state and prove some lemmas in what is called the full-information setting.

Two fundamental results in the adversarial, full-information setting

Setting: for $t = 1, 2, \ldots$

- the opponent picks losses $(\ell_{1t} \ldots \ell_{Kt}) \in [0,1]^K$

while, simultaneously,

- the decision-maker picks $I_t \in \{1 \ldots K\}$, possibly at random according to a distribution $p_t = (p_{1t} \ldots p_{Kt})$ and thanks to an auxiliary randomization U_t
- the decision-maker encurs a loss $\ell_{I_t t}$
- she gets to observe the full vector $(\ell_{1t} \ldots \ell_{Kt})$

Aim: control the regret $R_T = \sum_{t=1}^{T} \ell_{I_t, t} - \min_{k=1..K} \sum_{t=1}^{T} \ell_{kt}$

Algorithm: EWA (exponentially weighted averages) — parameter: $\eta > 0$

for $t \geq 1$,
for $k \in \{1..K\}$, $\quad p_{kt} = \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{ks}\right) \Big/ \sum_{i=1}^{K} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{is}\right)$

with the convention that an empty sum is null: so that $p_1 = (1/K, \ldots 1/K)$.

Theorem 1: For all opponents picking losses $\ell_{kt} \in [m, M]$, for all $T \geq 1$,

$$\widetilde{R}_T = \sum_{t=1}^{T} \sum_{i=1}^{K} p_{it} \ell_{it} - \min_{k=1..K} \sum_{t=1}^{T} \ell_{kt} \leq \frac{\ln K}{\eta} + \eta T \frac{(M-m)^2}{8}$$

In particular, $\widetilde{R}_T \leq (M-m)\sqrt{\frac{T}{2}\ln K}$ for $\eta = \frac{1}{M-m}\sqrt{\frac{8 \ln K}{T}}$

Practical use of the bound:

The Hoeffding-Azuma inequality ensures that for all T, with probability at least $1-\delta$, where $\delta > 0$:

$$R_T \leq \widetilde{R}_T + (M-m)\sqrt{\frac{T}{2}\ln\frac{1}{\delta}}$$

Proof: for a given t,

$$\sum_{i=1}^{K} p_{it}\,\ell_{it} \leqslant -\frac{1}{\eta}\ln\left(\sum_{i=1}^{K} p_{it}\,e^{-\eta \ell_{it}}\right) + \frac{\eta}{8}(M-m)^2$$

Indeed: Hoeffding's lemma ensures that if Z is a $[m,M]$-valued random variable, then

$$\ln \mathbb{E}\left[e^{-\eta Z}\right] \leqslant -\eta\,\mathbb{E}[Z] + \frac{\eta^2}{8}(M-m)^2$$

or put differently,

$$\mathbb{E}[Z] \leqslant -\frac{1}{\eta}\ln \mathbb{E}\left[e^{-\eta Z}\right] + \frac{\eta}{8}(M-m)^2$$

Thus,

$$\sum_{t=1}^{T}\sum_{i=1}^{K} p_{it}\,\ell_{it} \leqslant \eta\frac{T}{8}(M-m)^2 - \frac{1}{\eta}\sum_{t=1}^{T}\ln \underbrace{\sum_{i=1}^{K} p_{it}\,e^{-\eta \ell_{it}}}_{= \frac{\sum_{k=1}^{K}\exp\left(-\eta\sum_{s=1}^{t}\ell_{ks}\right)}{\sum_{k=1}^{K}\exp\left(-\eta\sum_{s=1}^{t-1}\ell_{ks}\right)}}$$

by def. of p_t

$$\underset{\text{telescoping sum}}{\leqslant} \eta T(M-m)^2/8 - \frac{1}{\eta}\ln\frac{\sum_{k=1}^{K}\exp\left(-\eta\sum_{t=1}^{T}\ell_{kt}\right)}{K}$$

The proof is concluded by rearranging the inequality and by resorting to the lower bound

$$\ln\left(\sum_{k=1}^{K}\exp\left(-\eta\sum_{t=1}^{T}\ell_{kt}\right)\right) \geqslant \ln\left(\max_{k=1..K}\exp\left(-\eta\sum_{t=1}^{T}\ell_{kt}\right)\right)$$

$$= \max_{k=1\ldots K} -\eta\sum_{t=1}^{T}\ell_{kt}$$

$$= -\eta\min_{k=1\ldots K}\sum_{t=1}^{T}\ell_{kt}$$

Our estimated pseudo-losses $\hat{\ell}_{kt} = \dfrac{1-g_{I_t,t}}{p_{kt}}\mathbb{1}_{\{I_t=k\}}$ are $\geqslant 0$ but can be arbitrarily large (as we don't impose any lower bound on the p_{kt}).

Thus, to study EWA in the bandit setting, we rather resort to the following result:

Theorem 2: For all opponents picking losses $\ell_{kt} \geqslant 0$, for all $T \geqslant 1$,

$$\tilde{R}_T = \sum_{t=1}^{T}\sum_{i=1}^{K} p_{it}\,\ell_{it} - \min_{k=1..K}\sum_{t=1}^{T}\ell_{kt}$$

$$\leqslant \frac{\ln K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{i=1}^{K} p_{it}\,\ell_{it}^2$$

Proof: $e^{-x} \leqslant 1-x+x^2/2$ for all $x \geqslant 0$

so that, for all $\eta > 0$ and all random variables $Z \geqslant 0$ a.s.,

$$\mathbb{E}\left[e^{-\eta Z}\right] \leqslant 1 - \eta\,\mathbb{E}[Z] + \frac{\eta^2}{2}\,\mathbb{E}[Z^2]$$

$\ln(1+u) \leqslant u$ for all $u > -1$

$$\ln\left(\mathbb{E}\left[e^{-\eta Z}\right]\right) \leqslant -\eta\,\mathbb{E}[Z] + \frac{\eta^2}{2}\,\mathbb{E}[Z^2]$$

and finally:
$$\mathbb{E}[Z] \leqslant -\frac{1}{\eta} \ln \mathbb{E}\left[e^{-\eta Z}\right] + \frac{\eta}{2}\,\mathbb{E}[Z^2]$$

In particular, for all $t \geqslant 1$,
$$\sum_{i=1}^{K} p_{it}\,\ell_{it} \leqslant -\frac{1}{\eta} \ln\left(\sum_{i=1}^{K} p_{it}\, e^{-\eta \ell_{it}}\right) + \frac{\eta}{2} \sum_{i=1}^{K} p_{it}\,\ell_{it}^2$$

from which the bound follows by summation over t and via the same manipulations as the ones performed at the end of the previous proof.

Back to EWA on estimated pseudo-losses $\hat{\ell}_{kt}$ in the bandit setting

Theorem 2 ensures that

$$\sum_{t=1}^{T} \sum_{i=1}^{K} p_{it} \hat{\ell}_{it} - \min_{k=1..K} \sum_{t=1}^{T} \hat{\ell}_{kt} \leqslant \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{j=1}^{K} p_{jt} \hat{\ell}_{jt}^{\,2}$$

Now,

$$\sum_{i=1}^{K} p_{it} \hat{\ell}_{it} = \sum_{i=1}^{K} \cancel{p_{it}} \frac{1-g_{I_t,t}}{\cancel{p_{it}}} \mathbb{1}_{\{I_t=i\}} = (1-g_{I_t,t}) \sum_{i=1}^{K} \mathbb{1}_{\{I_t=i\}} = 1-g_{I_t,t}$$

Similarly,

$$\sum_{j=1}^{K} p_{jt} \hat{\ell}_{jt}^{\,2} = \sum_{j=1}^{K} p_{jt} \frac{1-g_{I_t,t}}{p_{jt}^2} \mathbb{1}_{\{I_t=j\}} \leqslant \sum_{j=1}^{K} \frac{1}{p_{jt}} \mathbb{1}_{\{I_t=j\}}$$

Thus, we have

$$(*) \qquad \sum_{t=1}^{T} (1-g_{I_t,t}) - \min_{k=1..K} \sum_{t=1}^{T} \hat{\ell}_{kt} \leqslant \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{j=1}^{K} \frac{\mathbb{1}_{\{I_t=j\}}}{p_{jt}}$$

By the tower rule, $\mathbb{E}\big[\hat{\ell}_{kt}\big] = \mathbb{E}\Big[\mathbb{E}\big[\hat{\ell}_{kt} \,|\, \mathcal{F}_{t-1}\big]\Big] = \mathbb{E}\big[1-g_{kt}\big]$

and $\mathbb{E}\left[\dfrac{\mathbb{1}_{\{I_t=j\}}}{p_{jt}}\right] = \mathbb{E}\Big[\dfrac{1}{p_{jt}} \mathbb{E}\big[\underbrace{\mathbb{1}_{\{I_t=j\}} \,|\, \mathcal{F}_{t-1}}_{=p_{jt}}\big]\Big] = 1$

So that, by taking expectations $\mathbb{E}[\ ]$ in both sides of $(*)$ and using that $\mathbb{E}[\min \dots] \leqslant \min \mathbb{E}[\dots]$, we get

$$\mathbb{E}\Big[\sum_{t=1}^{T} (1-g_{I_t,t}) - \min_{k=1..K} \sum_{t=1}^{T} \hat{\ell}_{kt}\Big] \leqslant \frac{\ln K}{\eta} + \frac{\eta}{2} KT$$

$$\geqslant \mathbb{E}\Big[\sum_{t=1}^{T} (1-g_{I_t,t})\Big] - \min_{k=1..K} \mathbb{E}\Big[\sum_{t=1}^{T} (1-g_{kt})\Big]$$

$$= \max_{k=1..K} \mathbb{E}\Big[\sum_{t=1}^{T} g_{kt}\Big] - \mathbb{E}\Big[\sum_{t=1}^{T} g_{I_t,t}\Big]$$

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

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

Conclude 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 $N \geqslant 2$ and $x^{(1)} \dots x^{(N)}$ in $\mathcal{X}$, as well as $T \geqslant 1$

Then, at each round, they simultaneously pick

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

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

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:)

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

By definition of the exponentially weighted average strategy with learning rate $\eta > 0$,

$$\sum_{t=1}^{T} \sum_{j=1}^{N} p_{j,t} \underbrace{f(x^{(j)}, y_t)}_{\substack{\text{corresponds} \\ \text{to } \ell_{j,t} \in [0,M]}} - \min_{k=1,\dots N} \sum_{t=1}^{T} f(x^{(k)}, y_t)$$

$$\leq \underbrace{\frac{\ln N}{\eta} + \eta \frac{M^2}{8} T}_{= M\sqrt{\frac{T}{2}\ln N}}$$

given our choice for η

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

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

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

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

$f(\cdot, y)$ convex $\forall y$

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

$\sup \Sigma \leq \Sigma \sup$

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

4) $$\frac{1}{T} \sum_{t=1}^{T} f(x_t, y_t)$$

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

by $(*)$

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

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

5) In sections (1)-(4) 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 $T \longrightarrow +\infty$: $\inf_x \sup_y f(x,y) \leq \sup_y \min_k f(x^{(k)},y)$

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

$$\inf_x \sup_y f(x,y) \leqslant \underbrace{\inf_{N \geqslant 1} \inf_{\{x^{(1)} \ldots x^{(N)}\} \subset \mathcal{X}} \sup_{y \in \mathcal{Y}} \min_{k=1 \ldots N} f(x^{(k)},y)}$$

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

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

- $\mathcal{X}$ compact : $\forall \delta > 0, \quad \exists N$ and $x^{(1)} \ldots x^{(N)}$ s.t.

$$\mathcal{X} \subset \bigcup_{j=1}^{N} \mathcal{B}(x^{(j)}, \delta)$$

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

$$\forall x,y, \qquad f(x,y) \geqslant \min_{j=1 \ldots N} f(x^{(j)},y) - \varepsilon$$ by uniform continuity

Taking $\inf_x$ then $\sup_y$:

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

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

and we let $\varepsilon \searrow 0$

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

Sequential optimization for general convex sets.

We so far focused our attention on the simplex

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

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 $t = 1, 2, \ldots$

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

Algorithm: Online gradient descent with fixed learning rate $\eta > 0$

called $OGD(\eta)$

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

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

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

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 $OGD(\eta)$ 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:

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

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

diameter of $\mathcal{X}$: $\forall p,q \in \mathcal{X}$, $\sum_j (p_j - q_j)^2$
$D^2 = 2$

$$\leq \sum_j |p_j - q_j| \leq \sum_j (p_j + q_j) = 2$$

equality achieved if p, q 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: $(M - (-M))^2 \sqrt{\frac{T}{2} \ln N} = M \sqrt{8 T \ln N}$

The dependency in N is suboptimal for OGD.

 Lecture 6 — March 15, 2017 Lecture 6 — March 15, 2017

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

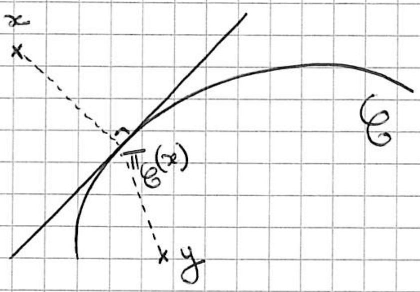
$\forall x \in \mathcal{C}, \quad \forall t, \qquad \ell_t(x_t) - \ell_t(x) \leqslant \nabla \ell_t(x_t) \cdot (x_t - x)$

Summing over t,

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

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

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



Facts: If $x \notin \mathcal{C}$, then

$\forall y \in \mathcal{C}, \qquad \big(y - \Pi_{\mathcal{C}}(x)\big) \cdot \big(x - \Pi_{\mathcal{C}}(x)\big) \leqslant 0$

Thus, $\forall y \in \mathcal{C}, \qquad \| y - \Pi_{\mathcal{C}}(x) \| \leqslant \| y - x \|.$

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

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

both are $\geqslant 0$

In particular, for $t \geqslant 1$:

$\forall x \in \mathcal{C},$

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

$$\underset{\text{by the facts above}}{\leqslant} \| x_t - \eta \nabla \ell_t(x_t) - x \|^2$$

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

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

$$\nabla \ell_t(x_t) \cdot (x_t - x) \leqslant \frac{1}{2\eta} \Big(\| x_t - x \|^2 - \| x_{t+1} - x \|^2 \Big) + \frac{\eta}{2} \underbrace{\| \nabla \ell_t(x_t) \|^2}_{\leqslant G^2}$$

by the assumption of bounded gradients

Summing over t:

$$\sum_{t=1}^{T} \nabla \ell_t(x_t) \cdot (x_t - x)$$

$$\leqslant \frac{1}{2\eta} \Big(\underbrace{\| x_1 - x \|^2}_{\leqslant D^2} - \underbrace{\| x_{T+1} - x \|^2}_{\geqslant 0} \Big) + \frac{\eta}{2} T G^2$$

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 3 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 $x_t = \Pi_{\mathcal{C}}\big(x_{t-1} - \eta_{t-1} \nabla \ell_{t-1}(x_{t-1})\big)$
  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
$$\eta_t = \begin{cases} 1 & \text{if } V_t = \displaystyle\sum_{s=1}^{t} \| \nabla \ell_s(x_s) \|^2 < D^2 \\ D/\sqrt{V_t} & \text{if } V_t \geq D^2 \end{cases}$$

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

without any prior knowledge of neither T nor G

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

Proof of the theorem and of the corollary:

Th: We extract from the proof of OGD(η) that

$$\forall x \in \mathcal{C}, \qquad \nabla \ell_t(x_t) \cdot (x_t - x) \leqslant \frac{1}{2\eta_t}\Big(\|x_t - x\|^2 - \|x_{t+1} - x\|^2 \Big) + \frac{\eta_t}{2} \|\nabla \ell_t(x_t)\|^2$$

Summing over t:

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

$$\leqslant \underbrace{\sum_{t=1}^T \frac{1}{2\eta_t} \Big(\|x_t - x\|^2 - \|x_{t+1} - x\|^2 \Big)} + \frac{1}{2} \sum_{t=1}^T \eta_t \|\nabla \ell_t(x_t)\|^2$$

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

Abel's transform:

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

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

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

Cor: Recall that we picked:

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

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

(Some practice actually explains why this is a natural choice $\rightarrow$ 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
$$\frac{D^2}{2\eta_T} + \frac{1}{2}\sum_{\substack{t:\\ V_t < D^2}} \|\nabla \ell_t(x_t)\|^2 + \frac{1}{2}\sum_{\substack{t:\\ V_t \geqslant D^2}} \frac{D}{\sqrt{V_t}} \|\nabla \ell_t(x_t)\|^2$$

We denote τ the largest $t \leqslant T$ such that $V_t < D^2$. Then, if $\tau < T$, we have $V_{\tau+1} \geqslant D^2$.

The above bound can be rewritten as:
$$\frac{D^2}{2\eta_T} + \underbrace{\frac{1}{2}\sum_{t: V_t < D^2} \|\nabla \ell_t(x_t)\|^2}_{= \frac{V_\tau}{2} < D^2/2} + \underbrace{\frac{1}{2} \underset{\substack{\uparrow \\ \text{possibly void sum} \\ (\text{if } \tau = T)}}{\sum_{t=\tau+1}^{T}} \frac{D}{\sqrt{V_t}} \|\nabla \ell_t(x_t)\|^2}$$

$$= \frac{1}{2}\sum_{t=\tau+1}^{T} \frac{D}{\sqrt{V_t}}\big(V_t - V_{t-1}\big) = \frac{1}{2}\sum_{t=\tau+1}^{T} \frac{D}{\sqrt{V_t}} \Big(\underbrace{\sqrt{V_t} + \sqrt{V_{t-1}}}_{\leqslant 2\sqrt{V_t}}\Big)\Big(\sqrt{V_t} - \sqrt{V_{t-1}}\Big)$$
$$\leqslant D \sum_{t=\tau+1}^{T} \Big(\sqrt{V_t} - \sqrt{V_{t-1}}\Big)$$
$$= D\Big(\sqrt{V_T} - \sqrt{V_\tau}\Big)$$

We conclude by distinguishing two cases:

- $\tau = T$: $\quad \eta_T = 1$ and the bound is $\leqslant \dfrac{D^2}{2} + \dfrac{D^2}{2} = D^2$
- $\tau < T$: $\quad \eta_T = D/\sqrt{V_T}$ and the bound is
$$\frac{D^2}{2\,D/\sqrt{V_T}} + \underbrace{\frac{V_\tau}{2}}_{\leqslant \frac{D\sqrt{V_\tau}}{2} \text{ as } V_\tau < D^2} + D\sqrt{V_T} - D\sqrt{V_\tau} \leqslant \frac{3}{2} D \sqrt{V_T}$$

↳ A readable and simple final bound is thus the one we indicated:
$$\max\Big\{D^2, \frac{3}{2}D\sqrt{V_T}\Big\} = \max\Big\{D^2, \; \frac{3}{2} D \sqrt{\sum_{t=1}^{T} \|\nabla \ell_t(x_t)\|^2}\Big\}$$