Fast rates with strongly convex loss functions

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 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 ?

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

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 $0 \le X_{(1)} \le X_{(2)} \le \dots \le X_{(N-1)} \le 1$ (order statistics) define N segments $(X_{(1)}, X_{(2)} - X_{(1)}, \dots, 1 - X_{(N-1)})$ 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 $\int_\chi g(p)\, d\mu(p) \simeq \frac{1}{m} \sum_{j=1}^{m} g(P_j)$
  - ↳ nonconvex 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:

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

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:

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

→ also, $\mu(\Delta_{\delta,\varepsilon}^*) = \varepsilon^{N-1}$ cf. properties of the Lebesgue measure (and ambient dimension $N-1$)

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

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

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

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...

Convex loss functions

Convex Loss Functions

$\ell_t : \mathcal{X} \to [m,M]$ 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

An inefficient algorithm to do so : the continuous EWA predictor

(↳ 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 $\ell_t : \mathcal{X} \to [m,M]$ of convex functions,

$$\sum_{t=1}^{T} \ell_t(p_t) - \inf_{q \in \mathcal{X}} \sum_{t=1}^{T} \ell_t(q) \leq \frac{\eta}{8}(M-m)^2 T + \inf_{\varepsilon \in \Delta(q_1)} \frac{-1}{\eta} \ln\left(\varepsilon^{N-1} e^{-\eta \varepsilon T (M-m)}\right)$$

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 ?

Convex functions & comparison to the best convex vector

↳ efficient forecaster

$\mathcal{X} = \{(p_1 \dots p_N) : \sum_j p_j = 1, \ \forall k \ p_k \geq 0\}$ is the simplex of convex weight vectors.

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

1. The statistician and the opponent pick simultaneously $p_t \in \mathcal{X}$ and a convex loss function $\ell_t : \mathcal{X} \to [m, M]$
2. ℓ_t and p_t are revealed

↳ Regret: $R_T = \sum_{t=1}^{T} \ell_t(p_t) - \inf_{p \in \mathcal{X}} \sum_{t=1}^{T} \ell_t(p)$ to be controled in a uniform way

Fact: Convex functions are subdifferentiable on the interior of their domain of definition:

Let $f : \mathcal{D} \to \mathbb{R}$ be convex, where $\mathcal{D} \subseteq \mathbb{R}^N$ is convex:

$\forall x \in \mathring{\mathcal{D}}, \ \exists \partial f_x \in \mathbb{R}^N \mid \ \forall y \in \mathcal{D}, \ f(x) - f(y) \leq \partial f_x \cdot (x - y)$

$\mathcal{G}_f(x) = \{$the set of all possible such $\partial f_x\}$ is called the subgradient of f at x

If f is differentiable at x, then $\mathcal{G}_f(x) = \{\nabla f(x)\}$.

Ex: $f(x) = |x|$, $\mathcal{G}_f(x) = \begin{cases} \{-1\} & \text{if } x < 0 \\ \{1\} & \text{if } x > 0 \\ [-1, 1] & \text{if } x = 0 \end{cases}$

Application: If $p_t \in \mathring{\mathcal{X}}$ (i.e. $p_{jt} > 0 \ \forall j$) then $\exists \partial \ell_t(p_t) \in \mathbb{R}^N \mid$

$\forall p \in \mathcal{X}, \ \ell_t(p_t) - \ell_t(p) \leq \partial \ell_t(p_t) \cdot (p_t - p)$

Example: Meta-statistical framework: $\ell_t(p) = \left(\sum_j p_j f_{jt} - y_t\right)^2$

ℓ_t is differentiable, $\nabla \ell_t(p) = \left(2\left(\sum_j p_j f_{jt} - y_t\right) f_{jt}\right)_j$

↳ Gradients in $[-G, G]^N$ with $G = 2B^2$ if $f_{jt}, y_t \in [0, B]$

Strategy: Exponentiated Gradients (EG) with learning rate $\eta > 0$

$p_1 = (1/N, \dots, 1/N)$ and denoting $\tilde{\ell}_{js} = j\text{-th component of } \partial \ell_s(p_s)$:

$$p_{jt} = \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{js}\right) \Big/ \sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{ks}\right)$$

($\tilde{\ell}_{js}$: interpreted as pseudo-loss)

Strategy easy to implement as soon as subgradients can be easily computed.

Theorem: Assume that the ℓ_t are picked such that

$$\forall t \ \forall j \ \forall p \in \mathcal{X} \qquad (\partial \ell_t(p))_j \in [-G, G]$$

Then the regret of EG tuned with $\eta > 0$ is controled as

$$R_T = \sum_{t=1}^{T} \ell_t(p_t) - \inf_{p \in \mathcal{X}} \sum_{t=1}^{T} \ell_t(p) \leq \frac{\ln N}{\eta} + \eta \frac{G^2 T}{2}$$

In particular, the choice $\eta = \frac{1}{G}\sqrt{\frac{2 \ln N}{T}}$ leads to $R_T \leq G\sqrt{2T \ln N}$.

Proof: Fix a p. Since by construction (cf. exponential weights) $p_t \in \mathcal{X}$, we have by convexity

$$\ell_t(p_t) - \ell_t(p) \leq \partial \ell_t(p_t) \cdot (p_t - p) \overset{\text{not.}}{=} \underbrace{\sum_{j=1}^{N} p_{jt} \tilde{\ell}_{jt}}_{} - \underbrace{\sum_{j=1}^{N} p_j \tilde{\ell}_{jt}}_{}$$

$\longrightarrow$ Reduction to linear losses: let's pretend that the opponent picks vectors of losses

$$(\tilde{\ell}_{1t}, \dots, \tilde{\ell}_{Nt}) \in [-G, G]$$

By the above inequality,

$$R_T \leq \sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \tilde{\ell}_{jt} - \inf_{p \in \mathcal{X}} \sum_{t=1}^{T} \sum_{j=1}^{N} p_j \tilde{\ell}_{jt}$$

$$= \sum_{t=1}^{T} \sum_{j=1}^{N} p_{jt} \tilde{\ell}_{jt} - \min_{k} \sum_{t=1}^{N} \tilde{\ell}_{kt} \leq \frac{\ln N}{\eta} + \eta \frac{(2G)^2}{8} T$$

↑ Bound proved last week in the linear case

This concludes the proof.

Application: Oracle inequalities → From individual sequences to stochastic sequences

Question: Let $(Y_1 \dots Y_T)$ be a sequence of iid random variables taking values in an arbitrary set $\mathcal{Y}$; and $Q : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a loss function, convex in its first argument, with bounded subgradients:

$$G = \sup_{y \in \mathcal{Y}} \sup_{p \in \mathcal{X}^*} \| \nabla Q(p, y) \|_\infty < +\infty$$

with either $Q \geq 0$ or $\forall \theta,\ Q(\theta, Y) \in L^1$ for all expectations below to be defined

Aim: Construct $\hat{\theta}_T = \hat{\theta}_T(Y_1 \dots Y_T)$ such that

$$\mathbb{E}\left[Q(\hat{\theta}_T, Y) \right] \leq \inf_{\theta \in \mathcal{X}} \mathbb{E}\left[Q(\theta, Y) \right] + \varepsilon_T$$

where Y is independent of the Y_t with the same distribution and $\varepsilon_T \searrow 0$.

(The expectation $\mathbb{E}$ is w.r.t. $Y_1 \dots Y_T$ and Y.)

Typical machine learning method: Empirical risk minimization, possibly in a regularized way:

$$\hat{\theta}_T \in \operatorname*{argmin}_{\theta \in \mathcal{X}} \left\{ \frac{1}{T} \sum_{t=1}^{T} Q(\theta, Y_t) + \lambda \operatorname{reg}(\theta) \right\}$$

with $\operatorname{reg}(\theta) = \|\theta\|_2$ or $\|\theta_1\|$ or ...

and λ to be tuned (called the regularization factor)

Our method: (as in our proof of Sion's lemma)

1) Pretend data is sequential, while of course it is batch

$$\tilde{\theta}_1 = (1/N, \dots 1/N)$$ and by induction, with $\eta = \frac{1}{G}\sqrt{\frac{2 \ln N}{T}}$:

$$\tilde{\theta}_{jt} = \exp\left(-\eta \sum_{s=1}^{t-1} \left(\nabla Q(\tilde{\theta}_s, Y_s) \right)_j \right) \Big/ \sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \left(\nabla Q(\tilde{\theta}_s, Y_s) \right)_k \right)$$

Strongly depends on the order picked

2) Consider an average: $\hat{\theta}_T = \frac{1}{T}\sum_{t=1}^{T} \tilde{\theta}_t$

(note: we do not use Y_t !)

Guarantee: $\mathcal{E}_T = G\sqrt{\frac{2\ln N}{T}}$

Proof: By the theorem on EG (exponentiated gradient) and given our choice of η:

$$\sum_{t=1}^{T} Q(\tilde{\theta}_t, Y_t) - \inf_{\theta \in \mathcal{X}} \sum_{t=1}^{T} Q(\theta, Y_t) \leq G\sqrt{2T\ln N}$$

Thus: $\forall \theta \in \mathcal{X}$, $\frac{1}{T}\sum_{t=1}^{T} Q(\tilde{\theta}_t, Y_t) \leq \frac{1}{T}\sum_{t=1}^{T} Q(\theta, Y_t) + \mathcal{E}_T$ with $\mathcal{E}_T = G\sqrt{\frac{2\ln N}{T}}$

Now, since $Y_1 \dots Y_T$ and Y are iid:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q(\theta, Y_t)\right] = \mathbb{E}[Q(\theta, Y)]$$

We conclude the proof by showing that $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q(\tilde{\theta}_t, Y_t)\right] \geq \mathbb{E}[Q(\hat{\theta}_T, Y)]$

Indeed: $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q(\tilde{\theta}_t, Y_t)\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q(\tilde{\theta}_t, Y)\right] \geq \mathbb{E}\left[Q\left(\frac{1}{T}\sum_t \tilde{\theta}_t, Y\right)\right]$

($=$: $\tilde{\theta}_t$ only depends on $Y_1 \dots Y_{t-1}$ and $(Y_1 \dots Y_{t-1}, Y_t)$ and $(Y_1 \dots Y_{t-1}, Y)$ have the same distribution)

($\geq$: by convexity and because the second argument is now the same for everyone!)

We have shown: $\forall \theta$, $\mathbb{E}[Q(\hat{\theta}_T, Y)] \leq \mathbb{E}[Q(\theta, Y)] + \mathcal{E}_T$

from which the conclusion follows.

Extension: Oracle inequality for stationary data.

A stationary sequence $(Y_1, Y_2, \dots)$ is by definition a sequence of random variables such that

$\forall k \geq 1$, $\forall t \geq 1$, $(Y_1 \dots Y_k)$ and $(Y_{t+1}, \dots Y_{t+k})$ have the same distribution.

The strategy is much more complicated to state and we need first a piece of notation

* For a sequence of observations $y_1, \dots y_t \in Y$, EG outputs the following weights, which we re-define as functions of the information available:

$$\psi_1^{EG} = \left(\tfrac{1}{N}, \dots \tfrac{1}{N}\right)$$

and for $t \geq 2$:

$$\psi_{jt}^{EG}(y_1, \dots y_{t-1}) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \left(\partial Q\left(\psi_s^{EG}(y_1 \dots y_{s-1}), y_s\right)\right)_j\right)}{\sum_{k=1}^{N} \exp\left(-\eta \sum_{s=1}^{t-1} \left(\partial Q\left(\psi_s^{EG}(y_1 \dots y_{s-1}), y_s\right)\right)_k\right)}$$

j-th component of $\psi_t^{EG}(y_1 \dots y_{t-1}) \in \mathbb{R}^N$

⚠ ψ_t^{EG} is an extremely complicated function of its arguments $y_1 \dots y_{t-1}$ (cf. the various calls to ψ_s^{EG}, with $s \leq t-1$, in its definition).

* We define $\tilde{\theta}_t = \psi_t^{EG}(Y_{T+2-t}, \dots Y_T)$ for $t \geq 2$ and $\tilde{\theta}_1 = \left(\tfrac{1}{N}, \dots \tfrac{1}{N}\right)$

We recommend: $\hat{\theta}_T = \frac{1}{T} \sum_{t=1}^{T} \tilde{\theta}_t$

⚠ Computing $\hat{\theta}_T$ is costly: the computation of each $\tilde{\theta}_t$ requires $t-1$ steps, namely the computation of the

$$\psi_s^{EG}\left(Y_{T+2-t}, \dots Y_{\underbrace{T-t}_{+s}}\right) \text{ for } 2 \leq s \leq t$$

no such intermediate quantity appears twice

$\Rightarrow$ about $\sum_{t=1}^{T} (t-1) = O(T^2/2)$ computation steps required.

What does EG guarantee ?

$$\forall \theta \in \mathcal{X}, \quad \frac{1}{T}\sum_{t=1}^{T} Q\left(\psi_t^{EG}(y_1, \dots y_{t-1}),\ y_t\right) \leq \frac{1}{T}\sum_{t=1}^{T} Q(\theta, y_t) + \varepsilon_T$$

still $G\sqrt{(2\ln N)/T}$

We take expectations in both sides to conclude :

- Stationarity implies in particular that all the y_t have the same distribution (but they are NOT independant in general); hence

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q(\theta, y_t)\right] = \mathbb{E}\left[Q(\theta, y_{T+1})\right]$$

- it also implies that $\left(\psi_t^{EG}(y_1, \dots y_{t-1}),\ y_t\right)$ and $\left(\psi_t^{EG}(y_{T+2-t}, \dots y_T),\ y_{T+1}\right) = (\tilde{\theta}_t, y_{T+1})$ (this defines $\tilde{\theta}_t$) have the same distribution, so that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q\left(\psi_t^{EG}(y_1, \dots y_{t-1}),\ y_t\right)\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} Q(\tilde{\theta}_t, y_{T+1})\right] \geq \mathbb{E}\left[Q(\hat{\theta}_T, y_{T+1})\right]$$

(convexity of $\theta \mapsto Q(\theta, y_{T+1})$)

- thus we proved, for our choice of $\hat{\theta}_T = \frac{1}{T}\sum_{t=1}^{T} \tilde{\theta}_t$

$$\mathbb{E}\left[Q(\hat{\theta}_T, y_{T+1})\right] \leq \inf_{\theta \in \mathcal{X}} \mathbb{E}\left[Q(\theta, y_{T+1})\right] + \varepsilon_T$$

same ε_T as for the iid case.

Sequential optimization for general convex sets

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 $\eta > 0$

- 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).

called $\mathrm{OGD}(\eta)$

$\|\cdot\|$ 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 at p, q Dirac masses $i \neq k$

- 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: $(M - (-M))\sqrt{\frac{T}{2} \ln N} = M\sqrt{2T \ln N}$

The dependency in N is suboptimal for OGD.

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

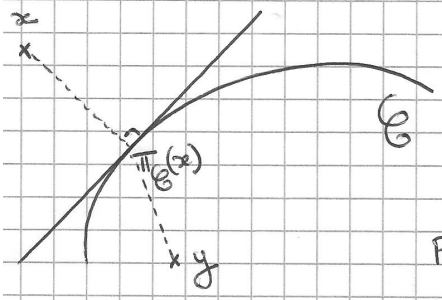
$$\forall x \in \mathcal{C}, \ \forall t, \quad \ell_t(x_t) - \ell_t(x) \leq \nabla \ell_t(x_t) \cdot (x_t - x)$$

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

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)

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

$$\leq \max\left\{D^2, \frac{3}{2} DG\sqrt{T}\right\}$$

without any prior knowledge of neither T nor G

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

Proof of the theorem and of the corollary:

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

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

Summing over t:

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

$$\leq \underbrace{\sum_{t=1}^T \frac{1}{2\eta_t}\left(\|x_t - x\|^2 - \|x_{t+1} - x\|^2 \right)}_{\text{We need only to show that this sum is } \leq D^2/2\eta_T} + \frac{1}{2}\sum_{t=1}^T \eta_t \|\nabla \ell_t(x_t)\|^2$$

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:

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

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

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

impossible in practice choice $\eta_t \equiv \eta = D/V_T$ 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 $V_t < D^2$ (possibly $t=0$ if $V_1 \geq D^2$). Then, if $\tau < T$, we have $V_{\tau+1} \geq D^2$. If $\tau = T$, then $\begin{cases} V_t < D^2 & \forall t \leq T \\ \eta_t = 1 \end{cases}$

The above bound can be rewritten as:

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

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

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

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

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

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

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

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