2) L6

Lecture 6: Online prediction with expert advice

Related to:
- bandits
- stochastic optimizat°
- classical statistics

References:
- "Predict°, learning and games", N. Cesa-Bianchi and G. Lugosi, Chapter 2.
- minimax desc → Survey: "Convex optim.: Algorithms and complexity" S. Bubeck
- Survey: "Regret analysis of sto. and non sto. multi-armed bandit pb" Bubeck and Cesa-Bianchi

I Setting
II Convex case
III Non convex case

I Setting

1.1 The problem

A statistician/learner wants to predict round after round the values of a sequence $y_1, y_2, y_3, \ldots \in Y$
His/her predict° are denoted by
$\hat{a}_1, \hat{a}_2, \ldots \in D$ ← not necessarily Y
↑ convex subset of a vector space

Sequential "aggregat°": at each round $t \geq 1$, the statist. has access to K base predict° ("the expert advice")
$a_{1,t}, \ldots, a_{K,t} \in D$

that can be combine ("aggregate") to choose his/her own predict° $\hat{a}_t \in \mathcal{D}$. (typically, $\hat{a}_t$ is a weighted average of the $a_{i,t}$)

Evaluat° of the predict° quality

loss funct° $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$

$(a, y) \mapsto \ell(a, y)$ measure how a "is close" to y

Goal of the statistician: on the long run the goal is to output predict° that are on average almost as good as those of the best expert in hindsight, ie, to minimize the "regret"

$$\mathrm{Reg}_T = \underbrace{\sum_{t=1}^{T} \ell(\hat{a}_t, y_t)}_{\text{cumulative loss of the statistician}} - \underbrace{\min_{1 \le i \le K} \sum_{t=1}^{T} \ell(a_{i,t}, y_t)}_{\text{cum. loss of the best expert in hindsight}}$$

Assumpt°: ? almost none!

We will prove guarantees of the form $\mathrm{Reg}_T \le \dots$ for all sequences $(y_t)_{t \ge 1}$ in $\mathcal{Y}$ and $(a_t)_{t \ge 1}$ in $\mathcal{D}^K$, $a_t = (a_{i,t})_{1 \le i \le K}$

provided: ℓ is convex in its first argument and bounded.

In particular: $y_t, t \ge 1$ not necessarily i.i.d.

All results stated below will be valid in the following scenarios:

- Meta-statistical setting: the sequence $(y_t)_{t \ge 1}$ is the realizat° of a stochastic process and the $a_{i,t}$ are

sequential estimations coming from K different statistical method.
→ the goal is to learn as fast as the best statistical method in our toolbox.

• deterministic setting
We make no stochastic assumpt° on $(y_t)_{t \geq 1}$
$a_{i,t}$ = predict° output by engineering / physical models.

Example: y_t = max amplitude of the ozone concentrat° at ground level (t: index of day)
$a_{i,t}$ = predict° made by chemico-physical models

• adversarial setting:
The y_t and the $a_{i,t}$ can be chosen by a malicious adversary who wants to defeat the statistician
Example: computer security, spam detect°

Formal recap' of the setting:
At every round $t \geq 1$:

1. the expert advice $a_{1,t}, \dots, a_{K,t} \in D$ are revealed to the statistician
2. the statistician makes her own predict° $\hat{a}_t \in D$, using the $a_{i,t}$ but also the past data $(y_s, a_s)_{1 \leq s \leq t-1}$
3. The statistician observes $y_t \in Y$ and suffers the loss $\ell(\hat{a}_t, y_t)$

NB: The above online protocol enables us to make out the dependencies between all quantities, eg, $\hat{a}_t = \hat{a}_t\left((y_s, a_s)_{1 \leq s \leq t-1}, a_t\right)$

A "strategy" is a sequence of funct° $(\hat{a}_t(\cdot))_{t\geq 1}$
where $\hat{a}_t : (Y \times \mathcal{D}^K)^{t-1} \times \mathcal{D}^K \to \mathcal{D}$

Goal: minimizing the regret

More precisely, the goal of this lecture is to build a strategy $(\hat{a}_t(\cdot))_{t\geq 1}$
s.t.

$$\sup_{\substack{(y_t)_{t\geq 1} \text{ in } Y \\ (a_t)_{t\geq 1} \text{ in } \mathcal{D}^K}} \left\{ \frac{1}{T}\sum_{t=1}^{T} \ell(\hat{a}_t, y_t) - \min_{1\leq i\leq K} \frac{1}{T}\sum_{t=1}^{T} \ell(a_i, y_t) \right\} \leq o(1)$$

as $T \to +\infty$

This implies that the average loss of the statistician is almost as small as that of the best expert.
This is true for all sequences $(y_t)_{t\geq 1}$ and $(a_t)_{t\geq 1}$
-> this theory is sometimes called "prediction of individual sequences", the theorems are valid for every seq.

Examples of $\mathcal{D}$, Y and ℓ

- Square loss for bounded prediction and observation
  $\mathcal{D} = Y = [-B; B]$; $\ell(a,y) = (a-y)^2$

- Absolute loss for binary classification
  $\mathcal{D} = [0,1]$, $Y = \{0,1\}$, $\ell(a,y) = |a-y|$

- Linear loss on the simplex
  $\mathcal{D} = \Delta(K) := \{v \in \mathbb{R}_+^K : \sum_{i=1}^{K} v_i = 1\}$
  probability vect. on K elements
  $Y = [0,1]^K$
  $\ell(a,y) = a \cdot y = \sum_{i=1}^{K} a_i y_i$

2/L6

1-2- Why is convex aggregat° useful?

Imagine the simple problem:

$$\mathcal{D} = \{0,1\}, \quad \mathcal{Y} = \{0,1\},$$

$$\ell(a,y) = |a-y| = \mathbb{1}_{\{a \neq y\}} \quad (\text{because } a, y \in \{0,1\})$$

At each round $t \geq 1$, $\begin{cases} a_{1,t} = 0 \\ a_{2,t} = 1 \end{cases}$

Important comment: whatever $\hat{a}_t \in \{0,1\}$ there exists $y_t \in \{0,1\}$ s.t. $\ell(\hat{a}_t, y_t) = 1$ (just take $y_t = 1 - \hat{a}_t$ so that $\sum_{t=1}^{T} \ell(\hat{a}_t, y_t) = T$

but $\min_{1 \leq i \leq 2} \sum_{t=1}^{T} \ell(a_{i,t}, y_t) \leq \frac{1}{2} \sum_{i=1}^{2} \sum_{t=1}^{T} \ell(a_{i,t}, y_t) = \frac{T}{2}$

As a consequence,

$$\frac{\text{Reg}_T}{T} = \frac{1}{T} \sum_{t=1}^{T} \ell(\hat{a}_t, y_t) - \min_{1 \leq i \leq 2} \frac{1}{T} \sum_{t=1}^{T} \ell(a_{i,t}, y_t)$$

$$\geq \frac{1}{2}, \quad \text{not} \leq o(1) !$$

The problem is that $\mathcal{D}$ is not convex.

If $\mathcal{D} = [0,1]$ instead, we could chose $\hat{a}_t \in [0,1]$ (eg $\frac{1}{2}$) and probably achieve $\text{Reg}_T \lesssim \sqrt{T}$ so that

$$\frac{\text{Reg}_T}{T} \lesssim \frac{1}{\sqrt{T}} = o(1)$$

(constants are omitted)

Take-home msg: convexity is crucial!

II The convex case

2-1 - The exponentially weighted average forecaster ("EWA")

Popular online algorithm / strategy dating back to the 90's -

EWA algorithm

- parameter : $\eta > 0$
- At each round $t \geq 1$,
  - * collect the expert advice $a_{i,t}$ $1 \leq i \leq K$
  - * compute the weight / proba vector $p_t \in \Delta(K)$ as

$$p_{i,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell(a_{i,s}, y_s)\right)}{\sum_{j=1}^{K} \exp\left(-\eta \sum_{s=1}^{t-1} \ell(a_{j,s}, y_s)\right)} \quad 1 \leq i \leq K$$

  - * compute the average prediction

$$\hat{a}_t = \sum_{i=1}^{K} p_{i,t}\, a_{i,t} \in D$$

(because $a_{i,t} \in D$ and D is convex)

  - * receive the observation $y_t \in Y$

Theorem: (Cesa-Bianchi 1999)

Let $T \geq 1$ and $K \geq 2$.

Assume that $\ell : D \times Y \to [B_1, B_2]$ is convex in its first argument, i.e. $a \mapsto \ell(a, y)$ is convex for all $y \in Y$.

Then, the EWA algorithm tuned with $\eta > 0$ satisfies, for all sequences $(y_t)_{t \geq 1}$ in Y and $(a_t)_{t \geq 1}$ in D^K,

$$\sum_{t=1}^{T} \ell(\hat{a}_t, g_t) + \min_{1\le i\le K} \sum_{t=1}^{T} \ell(a_{i,t}, g_t) \le \frac{\ln K}{\eta} + \frac{\eta T (B_2-B_1)^2}{8}$$

In particular, the choice of $\eta = \frac{1}{B_2-B_1}\sqrt{\frac{8\ln K}{T}}$ yields $\mathrm{Reg}_T \le (B_2-B_1)\sqrt{\frac{T\ln K}{2}}$

Consequence: for $\eta = \frac{1}{B_2-B_1}\sqrt{\frac{8\ln K}{T}}$, $\frac{\mathrm{Reg}_T}{T} \le (B_2-B_1)\sqrt{\frac{\ln K}{2T}} = O(1)$ $T\to+\infty$

Proof: we set

$$\ell_{i,t} = \ell(a_{i,t}, g_t)$$

$$L_{i,t} = \sum_{s=1}^{t} \ell_{i,s} \quad \text{(cumsum of expert } i\text{)}$$

$L_{i,0} = 0$ convent°

$$W_t = \frac{1}{K}\sum_{i=1}^{K} e^{-\eta L_{i,t-1}}$$

The idea here is to control the quantity $\ln W_{T+1} - \ln W_1$ in two different ways.

- Upper bound:

$$\ln W_{T+1} - \ln W_1 = \sum_{t=1}^{T} \ln\left(\frac{W_{t+1}}{W_t}\right) \quad \text{(telescopic sum)}$$

$$= \sum_{t=1}^{T} \ln\left(\frac{\frac{1}{K}\sum_{i=1}^{K} e^{-\eta L_{i,t}}}{\frac{1}{K}\sum_{i=1}^{K} e^{-\eta L_{i,t-1}}}\right)$$

$$= \sum_{t=1}^{T} \ln\left(\underbrace{\frac{\sum_{i=1}^{K} e^{-\eta L_{i,t-1}}}{\sum_{i=1}^{K} e^{-\eta L_{i,t-1}}}}_{p_{i,t}} e^{-\eta \ell_{i,t}}\right)$$

Next we provide an upper bound on $\ln\left(\sum_{i=1}^{K} p_{i,t} e^{-\eta \ell_{i,t}}\right)$ using Hoeffding's lemma.

Hoeffding's lemma

If $X \in [a,b]$ a.s., then $\forall \lambda \in \mathbb{R}$, $\ln \mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \frac{\lambda^2 (b-a)^2}{8}$

We define a r.v I_t s.t. $\mathbb{P}(I_t = i) = p_{i,t}$ and $X = \ell_{I_t,t}$ which is a r.v with values in $[B_1, B_2]$ ($\ell : \mathcal{D} \times \mathcal{Y} \to [B_1, B_2]$)

We use Hoeffding's lemma with X and $\lambda = -\eta$

$$\ln \mathbb{E}\left[e^{-\eta(X-\mathbb{E}[X])}\right] \leq \frac{\eta^2 (B_2-B_1)^2}{8}$$

but $\mathbb{E}[X] = \mathbb{E}\left[\ell_{I_t,t}\right] = \sum_{i=1}^{K} p_{i,t}\, \ell_{i,t} = p_t \cdot \ell_t$

$\ell_t = (\ell_{i,t})_{i \leq K} \in \mathbb{R}^K$

and $\ln \mathbb{E}\left[e^{-\eta(X-\mathbb{E}[X])}\right]$

$$= \ln\left(\sum_{i=1}^{K} p_{i,t}\, e^{-\eta(\ell_{i,t} - p_t \cdot \ell_t)}\right)$$

$$= \eta\, p_t \cdot \ell_t + \ln\left(\sum_{i=1}^{K} p_{i,t}\, e^{-\eta \ell_{i,t}}\right)$$

Therefore,

$$\ln\left(\sum_{i=1}^{K} p_{i,t}\, e^{-\eta \ell_{i,t}}\right) \leq -\eta\, p_t \cdot \ell_t + \frac{\eta^2 (B_2-B_1)^2}{8}$$

Finally, summing over $t = 1, \dots, T$

$$\ln W_{T+1} - \ln W_1 \leq -\eta \sum_{t=1}^{T} p_t \ell_t + \frac{T \eta^2 (B_2-B_1)^2}{8}$$

3/16

• Lower bound

$$\ln W_{T+1} - \ln W_1 = \ln W_{T+1} - 0 = \ln\left(\sum_{i=1}^{k} e^{-\eta L_{i,T}}\right) - \ln k$$

$$\geq \ln\left(\max_{1\leq i\leq k} e^{-\eta L_{i,T}}\right) - \ln k$$

$$= -\eta \min_{1\leq i\leq k} L_{i,T} - \ln k$$

• Collecting the upper and lower bounds.

$$-\eta \min_{1\leq i\leq k} L_{i,t} - \ln k \leq \ln W_{T+1} - \ln W_1$$

$$\leq -\eta \sum_{t=1}^{T} p_t \cdot \ell_t + \frac{T\eta^2 (B_2 - B_1)^2}{8}$$

Rearranging terms and dividing by η, we get:

$$(1) \quad \sum_{t=1}^{T} p_t \cdot \ell_t - \min_{1\leq i\leq k} \sum_{t=1}^{T} \ell_{i,t} \leq \frac{\ln k}{\eta} + \frac{T\eta (B_2 - B_1)^2}{8}$$

(1) → will be used later

$\ell(a_{i,t}, y_t)$

To conclude, note that

$$\ell(\hat{a}_t, y_t) \underset{\text{def of } \hat{a}_t}{=} \ell\left(\sum_{i=1}^{k} p_{i,t}\, a_{i,t}, y_t\right)$$

$$\leq \sum_{i=1}^{k} p_{i,t}\, \ell(a_{i,t}, y_t)$$

since $\ell(\cdot, y_t)$ is convex

$$= p_t \cdot \ell_t$$

□

Remarks:

- Important: how to tune the parameter η in practice? The values of T or $B_2 - B_1$ may be unknown in practice, so that $\eta = \frac{1}{B_2 - B_1}\sqrt{\frac{8 \ln K}{T}}$ may be unfeasible.

~~Idea: η_t with B_2, B_1 ~ empirical min and max (proof these similar)~~

→ if T is unknown, taking $\eta_t = \frac{1}{B_2 - B_1}\sqrt{\frac{8 \ln K}{t}}$ works as well ie, it yields a similar regret bound

→ if $B_2 - B_1$ is unknown, we can tune η_t as a funct° of the past data

~→ similar regret bound, without needing to know T or $B_2 - B_1$.

- For special loss functions ℓ (e.g., strongly convex), it is possible to prove better regret bounds, e.g., $\text{Reg}_T \leq \ln K$ instead of $\text{Reg}_T \leq \sqrt{T \ln K}$

2-2 Applicat° to $\nabla \ell_t$; links with gradient descent

Assume in this subsection that $\mathcal{D} = \Delta(K)$

$\ell : \Delta(K) \times \mathcal{Y} \to \mathbb{R}$ is differentiable in its 1st arg.

We write $\nabla \ell(a, y)$ for the gradient of $a \mapsto \ell(a, y)$ at point a.

$a_{i,t} = (0, 0, \dots, 0, \underset{i}{1}, 0, \dots, 0) \in \mathbb{R}^K$

$= e_i$

So, if $p_t \in \Delta(K)$ then $\ell\left(\sum_{i=1}^K p_{i,t} a_{i,t}, y_t\right) = \ell(p_t, y_t)$

(stated otherwise, a pt $p_t \in \Delta(K)$ can be seen as a convex combinat° of the corners e_i of $\Delta(K)$)

Algorithm ("EG": Exponential gradient)

- parameter $\eta > 0$
- $\forall t \geq 1$,

$$p_{i,t} := \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \nabla_i \ell(p_s, g_s)\right)}{\sum_{j=1}^{K} \exp\left(-\eta \sum_{s=1}^{t-1} \nabla_j \ell(p_s, g_s)\right)}, \quad 1 \leq i \leq K$$

where ∇_i is the i-th coordinate of the gradient.

Theorem (Cesa-Bianchi 1999)

Let $T \geq 1$ and $K \geq 2$

Assume $\ell : \Delta(K) \times Y \to \mathbb{R}$ is differentiable in 1st arg, and $\|\nabla \ell\|_\infty \leq G$, then the EG algorithm tuned with $\eta > 0$ satisfies, $\forall (g_t)_{t \geq 1}$ in Y.

$$\mathrm{Reg}_T^{\text{conv}} = \sum_{t=1}^{T} \ell(p_t, g_t) - \inf_{w \in \Delta(K)} \sum_{t=1}^{T} \ell(w, g_t)$$

$$\leq \frac{\ln K}{\eta} + \frac{T \eta G^2}{2}$$

In particular choosing $\eta = \frac{1}{G}\sqrt{\frac{2 \ln K}{T}}$ leads to

$$\mathrm{Reg}_T^{\text{conv}} \leq G\sqrt{2 T \ln K}$$

NB: here the goal was more ambitious: instead of targeting $\min_{1 \leq i \leq K} \sum_{t=1}^{T} \ell(e_i, g_t)$ (e_i = i-th vertex) we compete against the best convex combinat° of the e_i, $1 \leq i \leq K$, ie we target $\inf_{w \in \Delta(K)} \sum_{t=1}^{T} \ell(w, g_t)$

which is always $\leq$ (and sometimes $<$) than $\min_{1 \leq i \leq K} \sum_{t=1}^{T} \ell(e_i, g_t)$

(take w in $\{e_i : 1 \leq i \leq K\}$)

Proof: uses the previous proof!

By convexity of $w \mapsto \ell(w, y_t)$

$$\ell(p_t, y_t) - \ell(w, y_t) \leq \underbrace{\nabla \ell(p_t, y_t)}_{\ell_t} \cdot (p_t - w)$$

(We define $\ell_t := \nabla \ell(p_t, y_t)$)

$$\leq p_t \ell_t - w \cdot \ell_t$$

Therefore,

$$\sum_{t=1}^{T} \ell(p_t, y_t) - \inf_{w \in \Delta(K)} \sum_{t=1}^{T} \ell(w, y_t)$$

$$\leq \sup_{w \in \Delta(K)} \left\{ \sum_{t=1}^{T} p_t \cdot \ell_t - \sum_{t=1}^{T} w \cdot \ell_t \right\}$$

$$= \sum_{t=1}^{T} p_t \cdot \ell_t - \min_{w \in \Delta(K)} \left\{ w \cdot \sum_{t=1}^{T} \ell_t \right\}$$

$$= \sum_{t=1}^{T} p_t \cdot \ell_t - \min_{1 \leq i \leq K} \sum_{t=1}^{T} \ell_{i,t}$$

($w \mapsto w \cdot \sum_t \ell_t$ attains its max on corners of $\Delta(K)$ i.e. on e_i)

by (1)

$$\leq \frac{\ln K}{\eta} + \frac{\eta T (B_2 - B_1)^2}{8}$$

because, by definition of EG algo, $p_{i,t} \propto \exp\left(-\eta \sum_{s=1}^{t-1} \ell_{i,s}\right)$ corresponds to EWA algo used with

$\ell_{i,t} = \nabla_i \ell(p_t, y_t)$

$|\ell_{i,t}| = |\nabla_i \ell(p_t, y_t)| \leq \|\nabla \ell\|_\infty \leq G$

so $B_2 - B_1 \leq 2G$

$$\leq \frac{\ln K}{\eta} + \frac{\eta T G^2}{2}$$

□

Take-home msg: thanks to convexity and by using the gradients of ℓ, we were able to compete against a larger set.

Exercise: Generalize the algo and the thm when
- $\mathcal{D}$ is any convex set
- $a_{i,t} \in \mathcal{D}$ arbitrary

$$\cdot \text{Reg}_T^{\text{conv}} = \sum_{t=1}^{T} \ell(a_t^A, y_t) - \inf_{\nu \in \Delta(K)} \sum_{t=1}^{T} \ell\left(\sum_{i=1}^{K} \nu_i a_{i,t}, y_t\right)$$

Links with gradient descent

The EG algorithm can be equivalently rewritten as $\forall t \geq 1$,

(i) $q_t = (q_{i,t})_{1 \leq i \leq K} \in \mathbb{R}_+^K$ s.t.

$$q_{i,t} = p_{i,t-1} e^{-\eta \nabla_i \ell(p_{t-1}, y_{t-1})}$$

(ii) $p_t = \dfrac{q_t}{\sum_{i=1}^{K} q_{i,t}} \in \Delta(K)$

Note that:

(i) is a gradient descent applied to the logarithm:

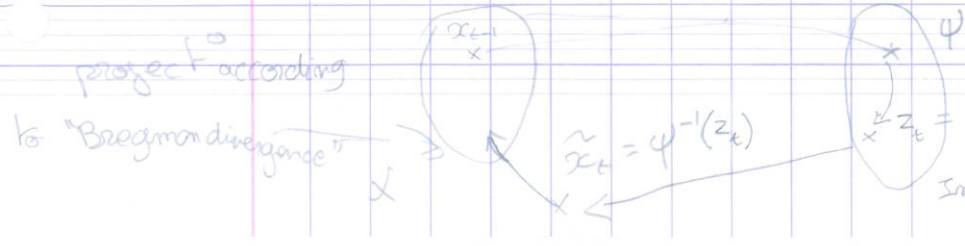
$$\ln(q_{i,t}) = \ln(p_{i,t-1}) - \eta \nabla_i \ell(p_{t-1}, y_{t-1})$$

(ii) is a project° step

$$q_t \in \underset{p \in \Delta(K)}{\operatorname{argmin}} \; KL(p, q_t) \quad \leftarrow \quad \sum_{i=1}^{K} p_i \ln\left(\frac{p_i}{q_{i,t}}\right)$$

Exercise ↑

Therefore, the EG algorithm is a special case of "mirror descent"



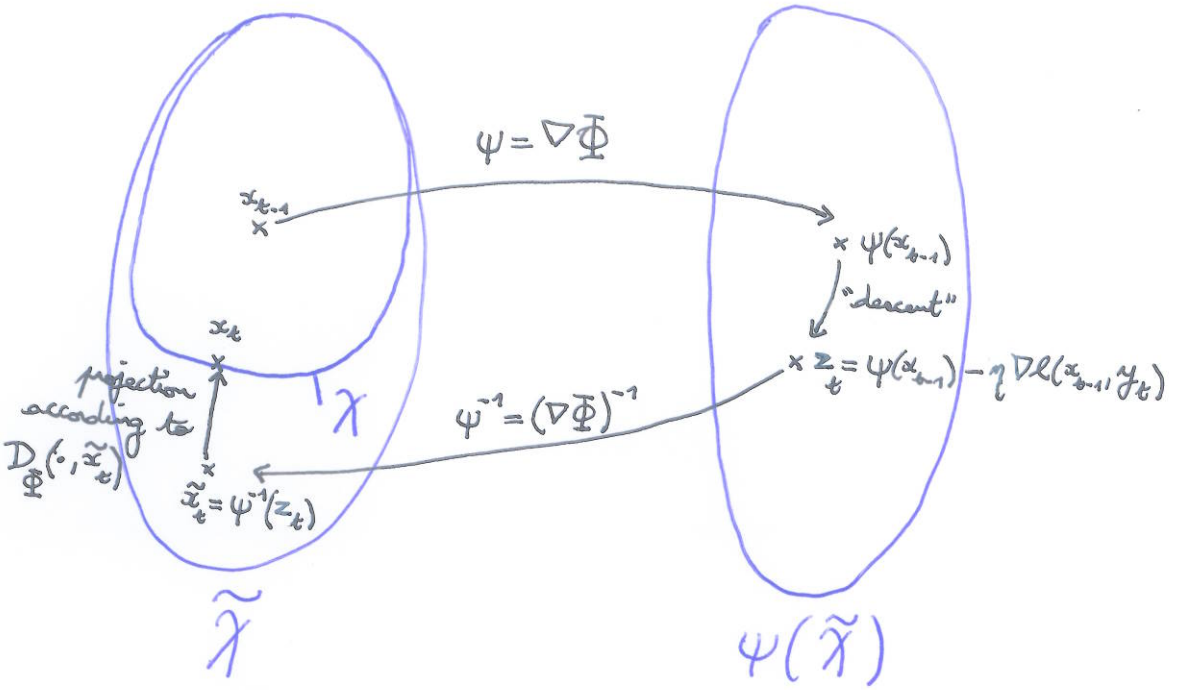


Fig: Online Mirror Descent

UNIVERSITÉ TOULOUSE III PAUL SABATIER
Université de Toulouse

In the above $\S$: $\mathcal{X} = \Delta(K)$

$\psi : \mathcal{X} \longrightarrow \mathbb{R}_+^K$

$(x_i)_{1 \le i \le K} \longmapsto (\ln(x_i))_{1 \le i \le K}$

- gradient step: $\ln \varphi_{t,\epsilon} = \ln \rho_{i,\epsilon} - \eta \nabla_x \ell(\rho_\epsilon, y_\epsilon)$
- $\psi : e^{\cdot}$

More generally: "online mirror descent"

Potential $\Phi : \tilde{\mathcal{X}} \to \mathbb{R}$ and convex + differentiable on $\mathring{\tilde{\mathcal{X}}}$ (interior)

$\mathcal{X} \subseteq \tilde{\mathcal{X}}$, $\mathcal{X}$: compact convex, $\tilde{\mathcal{X}}$: closed convex $\subseteq \mathbb{R}^K$

$\psi = \nabla \Phi$

Bregman divergence: the fct

$D_\Phi : \tilde{\mathcal{X}} \times \mathring{\tilde{\mathcal{X}}} \to \mathbb{R}$ (there are some assumptions on Φ near $\partial\tilde{\mathcal{X}}$)

$(y, x) \mapsto \Phi(y) - \Phi(x) - \nabla\Phi(x) \cdot (y - x)$

(difference between $\Phi(y)$ and its 1st order Taylor approx around x)

$\to \ge 0$ since Φ convex

Examples: 1) $\tilde{\mathcal{X}} = \mathbb{R}_+^{*K}$ $\quad \Phi(x) = \sum_{i=1}^K x_i (\ln x_i - 1)$

$D_\Phi(y, x) = \sum_{i=1}^K y_i \ln \frac{y_i}{x_i} - \sum_{i=1}^K (y_i - x_i)$; $\nabla\Phi(x) = (\ln x_i)_{1 \le i \le K}$

2) $\mathcal{X} = \mathbb{R}^K$ $\quad \Phi(x) = \frac{1}{2}\|x\|_2^2$, $\psi(x) = \nabla\Phi(x) = x$

gradient step here is the classical gradient step

$D_\Phi(y, x) = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|x\|_2^2 - x \cdot (y - x)$

$= \frac{1}{2}\|y - x\|_2^2$

Project° step: Euclidian project° onto $\mathcal{X}$.

Good news: $\exists$ theorem controlling the regret of mirror descent for many possible potential fct Φ.

II The non-convex case

See notes.

$\mathcal{D}$ is not assumed to be convex.

How can we choose $\hat{a}_t$ as a convex combinat° of the $a_{i,t}$? (doesn't exist!)

Instead, we can use a randomized algorithm:

$\forall t \geq 1$:

- compute $p_t \in \Delta(K)$ as in EWA
- instead of defining $\hat{a}_t = \sum_{i=1}^{K} p_{i,t} a_{i,t}$ we draw $I_t \sim p_t$ (conditionally on the past) and predict $\hat{a}_t = a_{I_t,t}$

In expectat°, everything is the same!

$$\mathbb{E}\left[\ell(a_{I_t,t}, y_t) \mid \underbrace{I_1, \dots, I_{t-1}}_{\text{past}} \right] = \sum_{i=1}^{K} p_{i,t} \ell(a_{i,t}, y_t) = p_t \cdot \ell_t$$

Therefore, the inequality (1)

$$\sum_{t=1}^{T} p_t \cdot \ell_t - \min_{1 \leq i \leq K} \sum_{t=1}^{T} \ell(a_{i,t}, y_t) \leq \frac{\ln K}{\eta} + \frac{T\eta (B_2 - B_1)^2}{8}$$

implies here that

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell(a_{I_t,t}, y_t) \right] - \min_{1 \leq i \leq K} \sum_{t=1}^{T} \ell(a_{i,t}, y_t) \leq \frac{\ln K}{\eta} + \frac{T\eta (B_2 - B_1)^2}{8}$$

assuming that $\ell : D \times Y \to [B_1, B_2]$ (no convexity) and if $a_{i,t}$ and y_t are deterministic.