Obtaining sparse and fast convergence rates online under Bernstein condition

Pierre Gaillard

September 19, 2017

INRIA Paris – SIERRA

This is joint work with Olivier Wintenberger

Introduction

SIERRA project team: from machine learning theory to applications.

- **Modelisation, prediction and control from training examples**
- **Algorithms:**
  - Scalability, stability and distribution
- **Theory:**
  - analysis of predictive and numerical performance
- **Applications through interdisciplinary collaborations**
  - Computer vision, bioinformatics, neuro-imaging, text, audio

We may start a collaboration with CWI in 2018.

January 2018: Dmitry Ostrovsky will join us as a INRIA/CWI postdoc

The setting of this talk: fit a sparse parameter online

Setting: Some data is sequentially observed. At each time step $t \geqslant 1$,

- a learner chooses a point $\widehat{\theta}_t \in \Theta \subset \mathbb{R}^d$ based on the past observations
- the environment then reveals a loss function $\ell_t : \theta \in \mathbb{R}^d \mapsto \mathbb{R}$ to evaluate the performance.

Goal: minimize our average error

$$\widehat{L}_n := \frac{1}{n} \sum_{t=1}^{n} \ell_t(\widehat{\theta}_t)$$

or fit the best parameter θ^* online if it exists.

An online setting? Because

- the data may only be sequentially observed: time-series (ex: weather forecast)
- nowadays the rate and volume of information flow are sharply increasing

The setting of this talk: fit a sparse parameter online

Setting: Some data is sequentially observed. At each time step $t \geqslant 1$,

- a learner chooses a point $\widehat{\theta}_t \in \Theta \subset \mathbb{R}^d$ based on the past observations
- the environment then reveals a loss function $\ell_t : \theta \in \mathbb{R}^d \mapsto \mathbb{R}$ to evaluate the performance.

Goal: minimize our average error

$$\widehat{L}_n := \frac{1}{n} \sum_{t=1}^{n} \left(Y_t - X_t \cdot \theta_t \right)^2$$

or fit the best parameter θ^* online if it exists.

An online setting? Because

- the data may only be sequentially observed: time-series (ex: weather forecast)
- nowadays the rate and volume of information flow are sharply increasing

The setting of this talk: fit a sparse parameter online

Setting: Some data is sequentially observed. At each time step $t \geqslant 1$,

- a learner chooses a point $\widehat{\theta}_t \in \Theta \subset \mathbb{R}^d$ based on the past observations
- the environment then reveals a loss function $\ell_t : \theta \in \mathbb{R}^d \mapsto \mathbb{R}$ to evaluate the performance.

Goal: minimize our average error

$$\widehat{L}_n := \frac{1}{n} \sum_{t=1}^{n} \ell_t(\widehat{\theta}_t)$$

or fit the best parameter θ^* online if it exists.

An online setting? Because

- the data may only be sequentially observed: time-series (ex: weather forecast)
- nowadays the rate and volume of information flow are sharply increasing

The setting of this talk: fit a sparse parameter online

Setting: Some data is sequentially observed. At each time step $t \geqslant 1$,

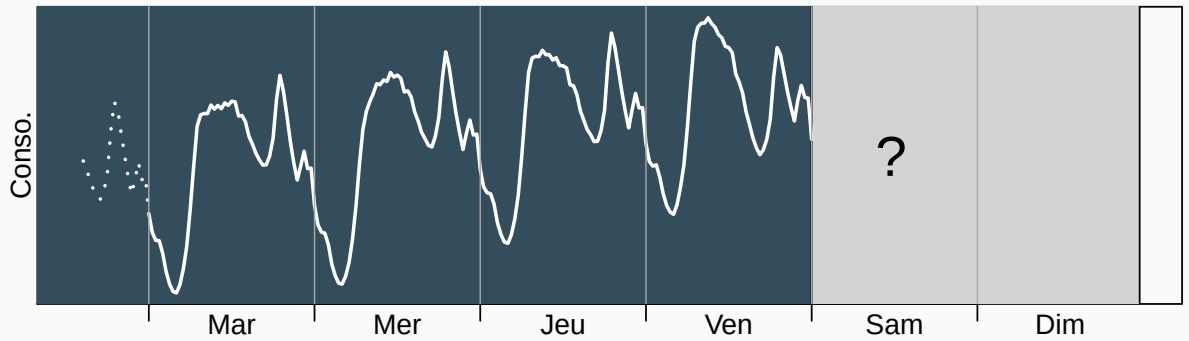
- a learner chooses a point $\widehat{\theta}_t \in \Theta \subset \mathbb{R}^d$ based on the past observations
- the environment then reveals a loss function $\ell_t : \theta \in \mathbb{R}^d \mapsto \mathbb{R}$ to evaluate the performance.

Goal: minimize our average error

$$\widehat{L}_n := \frac{1}{n} \sum_{t=1}^{n} \ell_t(\widehat{\theta}_t)$$

or fit the best parameter θ^* online if it exists.

Example: daily prediction of the electricity consumption



The setting of this talk: fit a sparse parameter online

Setting: Some data is sequentially observed. At each time step $t \geqslant 1$,

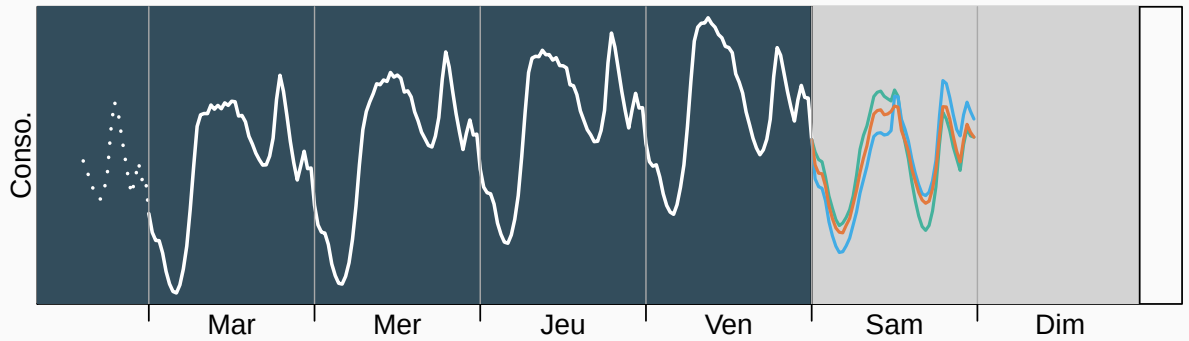
- a learner chooses a point $\widehat{\theta}_t \in \Theta \subset \mathbb{R}^d$ based on the past observations
- the environment then reveals a loss function $\ell_t : \theta \in \mathbb{R}^d \mapsto \mathbb{R}$ to evaluate the performance.

Goal: minimize our average error

$$\widehat{L}_n := \frac{1}{n} \sum_{t=1}^{n} \ell_t(\widehat{\theta}_t)$$

or fit the best parameter θ^* online if it exists.

Example: daily prediction of the electricity consumption



The setting of this talk: fit a sparse parameter online

Setting: Some data is sequentially observed. At each time step $t \geqslant 1$,

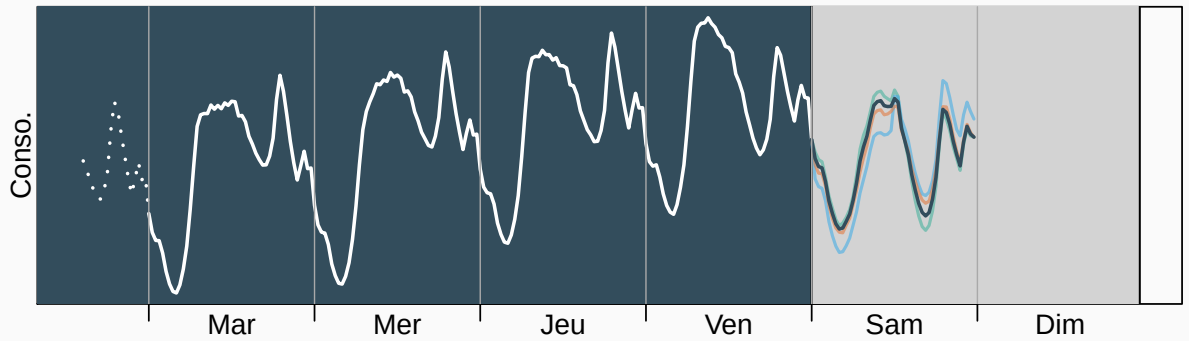
- a learner chooses a point $\widehat{\theta}_t \in \Theta \subset \mathbb{R}^d$ based on the past observations
- the environment then reveals a loss function $\ell_t : \theta \in \mathbb{R}^d \mapsto \mathbb{R}$ to evaluate the performance.

Goal: minimize our average error

$$\widehat{L}_n := \frac{1}{n} \sum_{t=1}^{n} \ell_t(\widehat{\theta}_t)$$

or fit the best parameter θ^* online if it exists.

Example: daily prediction of the electricity consumption



The setting of this talk: fit a sparse parameter online

Setting: Some data is sequentially observed. At each time step $t \geqslant 1$,

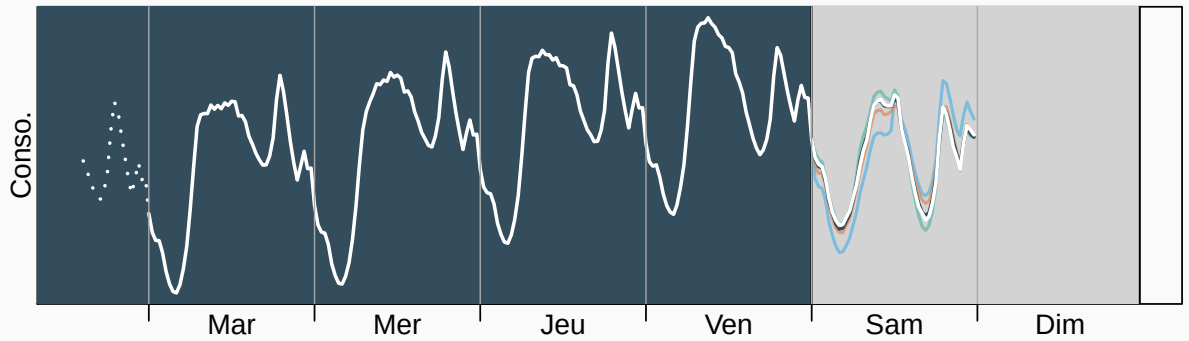
- a learner chooses a point $\widehat{\theta}_t \in \Theta \subset \mathbb{R}^d$ based on the past observations
- the environment then reveals a loss function $\ell_t : \theta \in \mathbb{R}^d \mapsto \mathbb{R}$ to evaluate the performance.

Goal: minimize our average error

$$\widehat{L}_n := \frac{1}{n} \sum_{t=1}^{n} \ell_t(\widehat{\theta}_t)$$

or fit the best parameter θ^* online if it exists.

Example: daily prediction of the electricity consumption



The setting of this talk: fit a sparse parameter online

Setting: Some data is sequentially observed. At each time step $t \geqslant 1$,

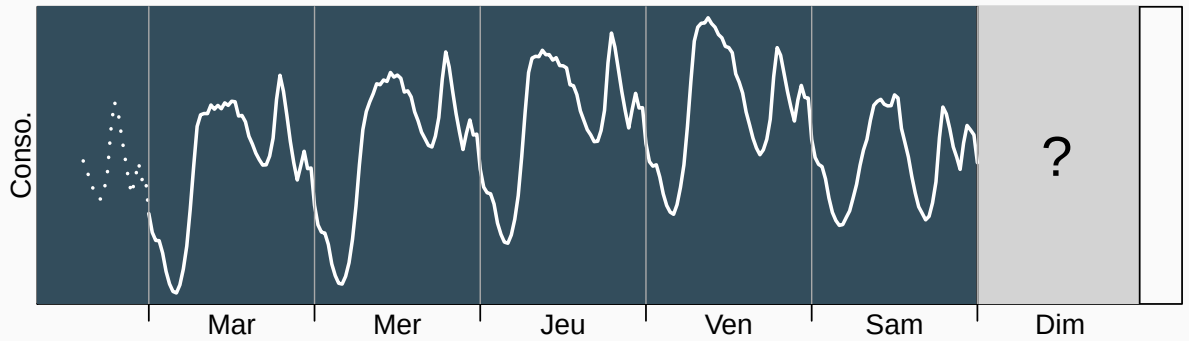
- a learner chooses a point $\widehat{\theta}_t \in \Theta \subset \mathbb{R}^d$ based on the past observations
- the environment then reveals a loss function $\ell_t : \theta \in \mathbb{R}^d \mapsto \mathbb{R}$ to evaluate the performance.

Goal: minimize our average error

$$\widehat{L}_n := \frac{1}{n} \sum_{t=1}^{n} \ell_t(\widehat{\theta}_t)$$

or fit the best parameter θ^* online if it exists.

Example: daily prediction of the electricity consumption



The setting of this talk: fit a sparse parameter online

Setting: Some data is sequentially observed. At each time step $t \geqslant 1$,

- a learner chooses a point $\widehat{\theta}_t \in \Theta \subset \mathbb{R}^d$ based on the past observations
- the environment then reveals a loss function $\ell_t : \theta \in \mathbb{R}^d \mapsto \mathbb{R}$ to evaluate the performance.

Goal: minimize our average error

$$\widehat{L}_n := \frac{1}{n} \sum_{t=1}^{n} \ell_t(\widehat{\theta}_t)$$

or fit the best parameter θ^* online if it exists.

In this talk, I will focus on cases where θ^*

- exists (nice i.i.d. losses)
- is of large dimension
- is sparse

Question: what guarantees can we ensure on $\widehat{L}_n$?

The regret: a relative criterion

If all parameters $\theta \in \Theta$ perform badly $\rightarrow$ hard to get a small error!

We evaluate our performance relatively

$$\underbrace{\frac{1}{n}\sum_{t=1}^{n} \ell_t(\widehat{\theta}_t)}_{\overset{\text{def}}{=} \widehat{L}_n} = \underbrace{\min_{\theta\in\Theta} \frac{1}{n}\sum_{t=1}^{n} \ell_t(\theta)}_{\overset{\text{def}}{=} L_n(\theta^\star)} + \underbrace{\frac{1}{n}\sum_{t=1}^{n} \ell_t(\widehat{\theta}_t) - \min_{\theta\in\Theta} \frac{1}{n}\sum_{t=1}^{n} \ell_t(\theta)}_{\overset{\text{def}}{=} \mathrm{Reg}_n}$$

our performance

reference performance (approximation error)

average regret (estimation error)

Goal

Perform almost as good as the best parameter θ when $n \rightarrow \infty$

$$\limsup_{n\rightarrow\infty} \left(\sup_{\text{data}} \mathrm{Reg}_n \right) \leqslant 0$$

Finite set $\Theta = \{\theta_1, \dots, \theta_K\}$

Algorithm: Hedge(η)[1]

At time $t \geqslant 1$, assign a weight to each $\theta_k \in \Theta$ based on its past performance

$$p_{k,t} = \frac{e^{-\eta \sum_{s=1}^{t-1} \ell_s(\theta_k)}}{\sum_{j=1}^{K} e^{-\eta \sum_{s=1}^{t-1} \ell_s(\theta_j)}}$$

and predict the combination: $\widehat{\theta}_t = \sum_{k=1}^{K} p_{k,t} \theta_k$

Performance guarantee: for **convex losses** ℓ_t and optimal learning rate $\eta > 0$

$$\mathrm{Reg}_n \overset{\mathrm{def}}{=} \frac{1}{n} \sum_{t=1}^{n} \ell_t(\widehat{\theta}_t) - \min_{1 \leqslant k \leqslant K} \frac{1}{n} \sum_{t=1}^{n} \ell_t(\theta_k) \lesssim \sqrt{\frac{\log K}{n}}$$

- We approach the best performance at the rate $n^{-1/2}$.
- This result holds without any stochastic assumption.
- $1/\sqrt{n}$ is slow.

[1] Littlestone and Warmuth, "The weighted majority algorithm", 1994; Vovk, "Aggregating strategies", 1990.

Finite set $\Theta = \{\theta_1, \dots, \theta_K\}$

Algorithm: Hedge(η)[1]

At time $t \geqslant 1$, assign a weight to each $\theta_k \in \Theta$ based on its past performance

$$p_{k,t} = \frac{e^{-\eta \sum_{s=1}^{t-1} \ell_s(\theta_k)}}{\sum_{j=1}^{K} e^{-\eta \sum_{s=1}^{t-1} \ell_s(\theta_j)}}$$

and predict the combination: $\widehat{\theta}_t = \sum_{k=1}^{K} p_{k,t} \theta_k$

Performance guarantee: for **convex losses** ℓ_t and optimal learning rate $\eta > 0$

$$\mathrm{Reg}_n \overset{\mathrm{def}}{=} \frac{1}{n} \sum_{t=1}^{n} \ell_t(\widehat{\theta}_t) - \min_{1 \leqslant k \leqslant K} \frac{1}{n} \sum_{t=1}^{n} \ell_t(\theta_k) \lesssim \sqrt{\frac{\log K}{n}}$$

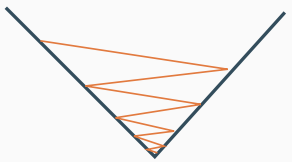
Can we do better than $1/\sqrt{n}$?

- No in general: optimal rate
- Yes under additional assumption: i.i.d. data, strongly convex loss,...

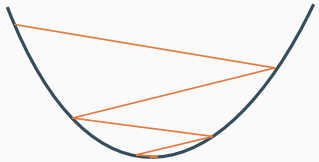
[1] Littlestone and Warmuth, "The weighted majority algorithm", 1994; Vovk, "Aggregating strategies", 1990.

Better rates than $1/\sqrt{n}$?

Yes, under strong convexity: it exits $\mu > 0$ such that $\theta \mapsto \ell_t(\theta) - \mu\|\theta\|^2$ is convex.



Convex

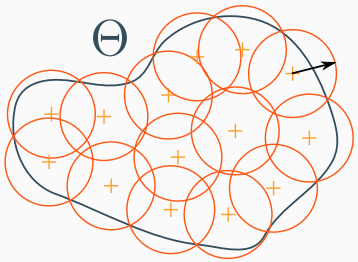


Strongly Convex

The average error is bounded as:

$$\widehat{L}_n \lesssim \min_{1 \leqslant k \leqslant K} \frac{1}{n} \sum_{t=1}^{n} \ell_t(\theta_k) + \begin{cases} \sqrt{\frac{\log K}{n}} & \text{Convex loss} \\ \frac{\log K}{\mu n} & \text{Strongly convex loss} \end{cases}$$

First solution for Lipschitz loss:
discretize + apply method for finite $\Theta = \{\theta_1, \dots, \theta_K\}$



We can decompose our error into three terms:

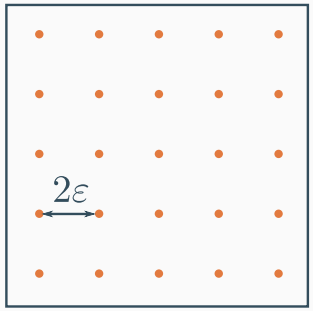
$$\widehat{L}_n = \underbrace{\min_{\theta \in \Theta} \frac{1}{n} \sum_{t=1}^{n} \ell_t(\theta)}_{\text{Best error in } \Theta} + \underbrace{\widehat{L}_n - \min_{1 \leqslant k \leqslant K} \frac{1}{n} \sum_{t=1}^{n} \ell_t(\theta_k)}_{\text{Regret with respect to finite set}} + \underbrace{\min_{1 \leqslant k \leqslant K} \frac{1}{n} \sum_{t=1}^{n} \ell_t(\theta_k) - \min_{\theta \in \Theta} \frac{1}{n} \sum_{t=1}^{n} \ell_t(\theta)}_{\text{Discretization error}}$$

$$\lesssim \min_{\theta \in \Theta} \frac{1}{n} \sum_{t=1}^{n} \ell_t(\theta) + \frac{\log \mathcal{N}(\Theta, \varepsilon)}{\mu n} + \varepsilon \qquad \leftarrow \text{ for strongly-convex loss}$$

where $\mathcal{N}(\Theta, \varepsilon)$ is the number of points needed to cover Θ at approximation ε

Example: compact ball in $\mathbb{R}^d$

If Θ is the L_∞ unit ball in $\mathbb{R}^d$ then $\mathcal{N}(\Theta, \varepsilon) \approx \left(\frac{1}{\varepsilon}\right)^d$



$\left(\frac{1}{\varepsilon}\right)^2$ are needed to cover $[-1,1]^2$.

Example: If $\Theta = [-1,1]^d$ and if the losses are strongly convex then

$$\text{Average regret} \lesssim \frac{\log \mathcal{N}(\Theta, \varepsilon)}{n} + \varepsilon \lesssim \frac{d \log(1/\varepsilon)}{\mu n} + \varepsilon \overset{\varepsilon = \frac{1}{n}}{\approx} \frac{d \log n}{\mu n}$$

👍 optimal rate (up to log) for strongly-convex function without any stochastic assumption

👎 inefficient

Solutions in the literature: Gradient descent, Online Newton Step,...

Our solution: build an adaptive discretization of Θ + gradient descent $\rightarrow$ better guarantees under sparsity

First step: linearize the loss using the gradient

Apply Hedge to a linearized and regularized loss $\tilde{\ell}_t(\theta) = \nabla \ell_t(\widehat{\theta}_t) \cdot \theta$ +Regularization

Algorithm: Hedge with regularized gradient losses

At time $t \geqslant 1$, assign a weight to each $\theta_k \in \Theta$ based on its past performance

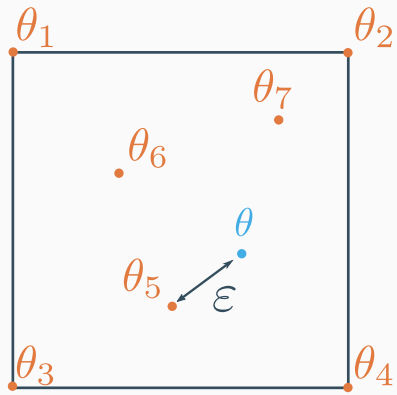
$$p_{k,t} = \frac{e^{-\eta \sum_{s=1}^{t-1} \tilde{\ell}_s(\theta_k)}}{\sum_{j=1}^{K} e^{-\eta \sum_{s=1}^{t-1} \tilde{\ell}_s(\theta_j)}}$$

and predict the combination: $\widehat{\theta}_t = \sum_{k=1}^{K} p_{k,t} \theta_k$

Performance: for convex losses ℓ_t, optimal learning rate $\eta > 0$, applying the above algorithm to a discretization set which contains the 2d corners ensures that for any $\theta \in [-1, 1]^d$

$$\frac{1}{n} \sum_{t=1}^{n} \ell_t(\widehat{\theta}_t) - \frac{1}{n} \sum_{t=1}^{n} \ell_t(\theta) \lesssim \varepsilon \sqrt{\frac{\log(K)}{n}} + \frac{\log(K)}{n}$$

where $\varepsilon = \min_{1 \leqslant k \leqslant K} \|\theta - \theta_k\|_1$.



- With only $2d$ points, we can approach the best parameter of $[-1, 1]^d$ at rate $1/\sqrt{n}$.
- $1/\sqrt{n}$ is slow: cannot be improved directly if ℓ_t are strongly convex
- small if ε is small

Second step: enlarge the discretization grid step by step

Assumptions

- **i.i.d. data:** $\ell_1, \ell_2, \dots$ are i.i.d., $\theta^* \in \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[\ell_t](\theta)$
- **Strong convexity:** for all $\theta \in \mathbb{R}^d$ $\mu \|\theta_1 - \theta^*\|_2^2 \leqslant \mathbb{E}[\ell_t(\theta_1) - \ell_t(\theta^*)]$

Then, if $\bar{\theta}_n := \frac{1}{n} \sum_{t=1}^n \widehat{\theta}_t$ w.h.p.

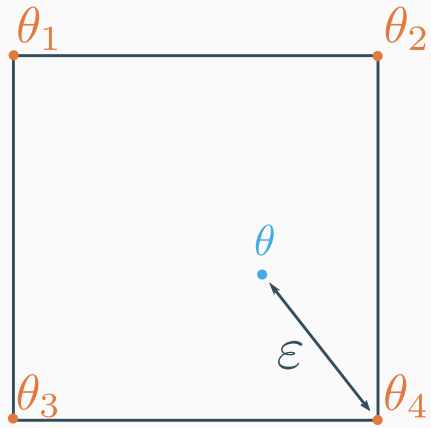
$$\left\| \bar{\theta}_n - \theta^* \right\|_1^2 \lesssim \frac{d}{\mu} \mathrm{Reg}_n$$

i.e. small regret $\Rightarrow$ w.h.p. $\left\| \bar{\theta}_n - \theta^* \right\|_1$ is small

The previous procedure ensures a regret

$$\mathrm{Reg}_n \lesssim \varepsilon \sqrt{\frac{\log d}{n}} + \frac{\log d}{n}$$

as soon as one θ_k is ε-close to θ^* in L1-norm.



Build an adaptive discretization

Enlarge the discretization set $\{\theta_1, \dots, \theta_K\}$ by sequentially adding the new estimators $\bar{\theta}_t$
$\rightarrow$ **auto-regulated procedure**

Second step: enlarge the discretization grid step by step

Assumptions

- **i.i.d. data:** $\ell_1, \ell_2, \dots$ are i.i.d., $\theta^* \in \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[\ell_t](\theta)$
- **Strong convexity:** for all $\theta \in \mathbb{R}^d$ $\mu \|\theta_1 - \theta^*\|_2^2 \leqslant \mathbb{E}[\ell_t(\theta_1) - \ell_t(\theta^*)]$

Then, if $\bar{\theta}_n := \frac{1}{n} \sum_{t=1}^n \widehat{\theta}_t$ w.h.p.

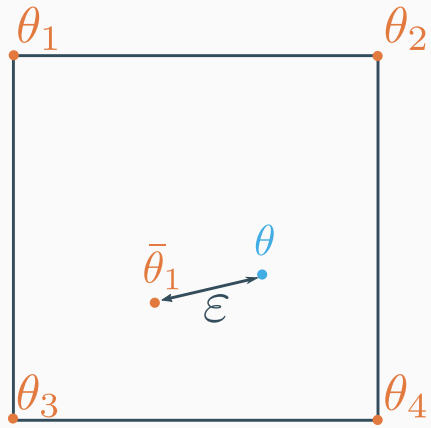
$$\left\| \bar{\theta}_n - \theta^* \right\|_1^2 \lesssim \frac{d}{\mu} \mathrm{Reg}_n$$

i.e. small regret $\Rightarrow$ w.h.p. $\left\| \bar{\theta}_n - \theta^* \right\|_1$ is small

The previous procedure ensures a regret

$$\mathrm{Reg}_n \lesssim \varepsilon \sqrt{\frac{\log d}{n}} + \frac{\log d}{n}$$

as soon as one θ_k is ε-close to θ^* in L1-norm.



Build an adaptive discretization

Enlarge the discretization set $\{\theta_1, \dots, \theta_K\}$ by sequentially adding the new estimators $\bar{\theta}_t$
$\rightarrow$ **auto-regulated procedure**

Second step: enlarge the discretization grid step by step

Assumptions

- **i.i.d. data:** $\ell_1, \ell_2, \dots$ are i.i.d., $\theta^* \in \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[\ell_t](\theta)$
- **Strong convexity:** for all $\theta \in \mathbb{R}^d$ $\mu \|\theta_1 - \theta^*\|_2^2 \leqslant \mathbb{E}[\ell_t(\theta_1) - \ell_t(\theta^*)]$

Then, if $\bar{\theta}_n := \frac{1}{n} \sum_{t=1}^n \widehat{\theta}_t$ w.h.p.

$$\left\| \bar{\theta}_n - \theta^* \right\|_1^2 \lesssim \frac{d}{\mu} \mathrm{Reg}_n$$

i.e. small regret $\Rightarrow$ w.h.p. $\left\| \bar{\theta}_n - \theta^* \right\|_1$ is small

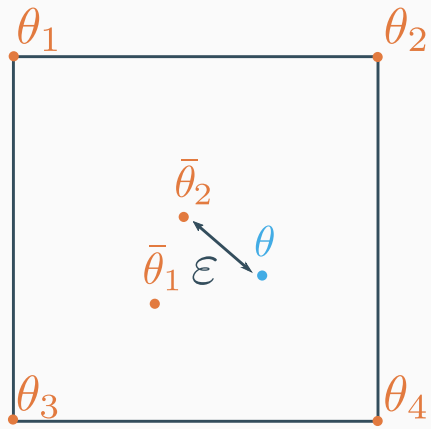
The previous procedure ensures a regret

$$\mathrm{Reg}_n \lesssim \varepsilon \sqrt{\frac{\log d}{n}} + \frac{\log d}{n}$$

as soon as one θ_k is ε-close to θ^* in L1-norm.

Build an adaptive discretization

Enlarge the discretization set $\{\theta_1, \dots, \theta_K\}$ by sequentially adding the new estimators $\bar{\theta}_t$
$\rightarrow$ **auto-regulated procedure**



Second step: enlarge the discretization grid step by step

Assumptions

- **i.i.d. data:** $\ell_1, \ell_2, \dots$ are i.i.d., $\theta^* \in \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[\ell_t](\theta)$
- **Strong convexity:** for all $\theta \in \mathbb{R}^d$ $\mu \|\theta_1 - \theta^*\|_2^2 \leqslant \mathbb{E}[\ell_t(\theta_1) - \ell_t(\theta^*)]$

Then, if $\bar{\theta}_n := \frac{1}{n} \sum_{t=1}^n \widehat{\theta}_t$ w.h.p.

$$\left\| \bar{\theta}_n - \theta^* \right\|_1^2 \lesssim \frac{d}{\mu} \mathrm{Reg}_n$$

i.e. small regret $\Rightarrow$ w.h.p. $\left\| \bar{\theta}_n - \theta^* \right\|_1$ is small

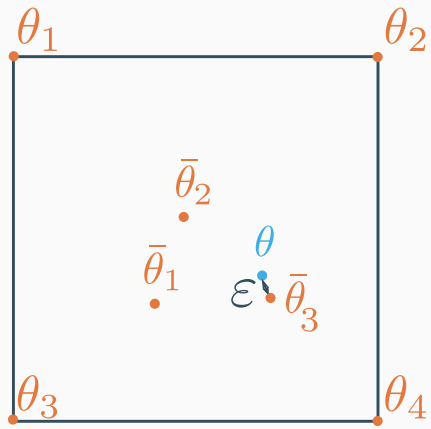
The previous procedure ensures a regret

$$\mathrm{Reg}_n \lesssim \varepsilon \sqrt{\frac{\log d}{n}} + \frac{\log d}{n}$$

as soon as one θ_k is ε-close to θ^* in L1-norm.

Build an adaptive discretization

Enlarge the discretization set $\{\theta_1, \dots, \theta_K\}$ by sequentially adding the new estimators $\bar{\theta}_t$
$\rightarrow$ **auto-regulated procedure**



Second step: enlarge the discretization grid step by step

Assumptions

- **i.i.d. data:** $\ell_1, \ell_2, \dots$ are i.i.d., $\theta^* \in \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[\ell_t](\theta)$
- **Strong convexity:** for all $\theta \in \mathbb{R}^d$ $\mu \|\theta_1 - \theta^*\|_2^2 \leqslant \mathbb{E}[\ell_t(\theta_1) - \ell_t(\theta^*)]$

Then, if $\bar{\theta}_n := \frac{1}{n} \sum_{t=1}^n \widehat{\theta}_t$ w.h.p.

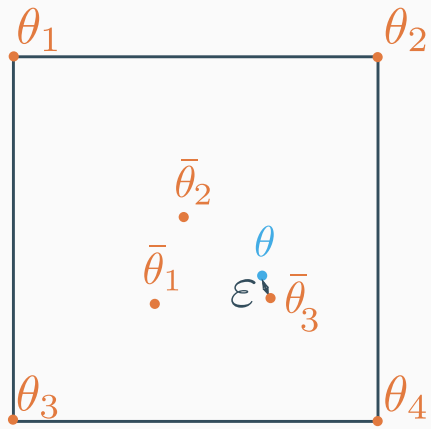
$$\left\| \bar{\theta}_n - \theta^* \right\|_1^2 \lesssim \frac{d}{\mu} \mathrm{Reg}_n$$

i.e. small regret $\Rightarrow$ w.h.p. $\left\| \bar{\theta}_n - \theta^* \right\|_1$ is small

The previous procedure ensures a regret

$$\mathrm{Reg}_n \lesssim \varepsilon \sqrt{\frac{\log d}{n}} + \frac{\log d}{n} \overset{\text{Fixed point}}{\lesssim} \frac{d}{\mu} \frac{\log d}{n}$$

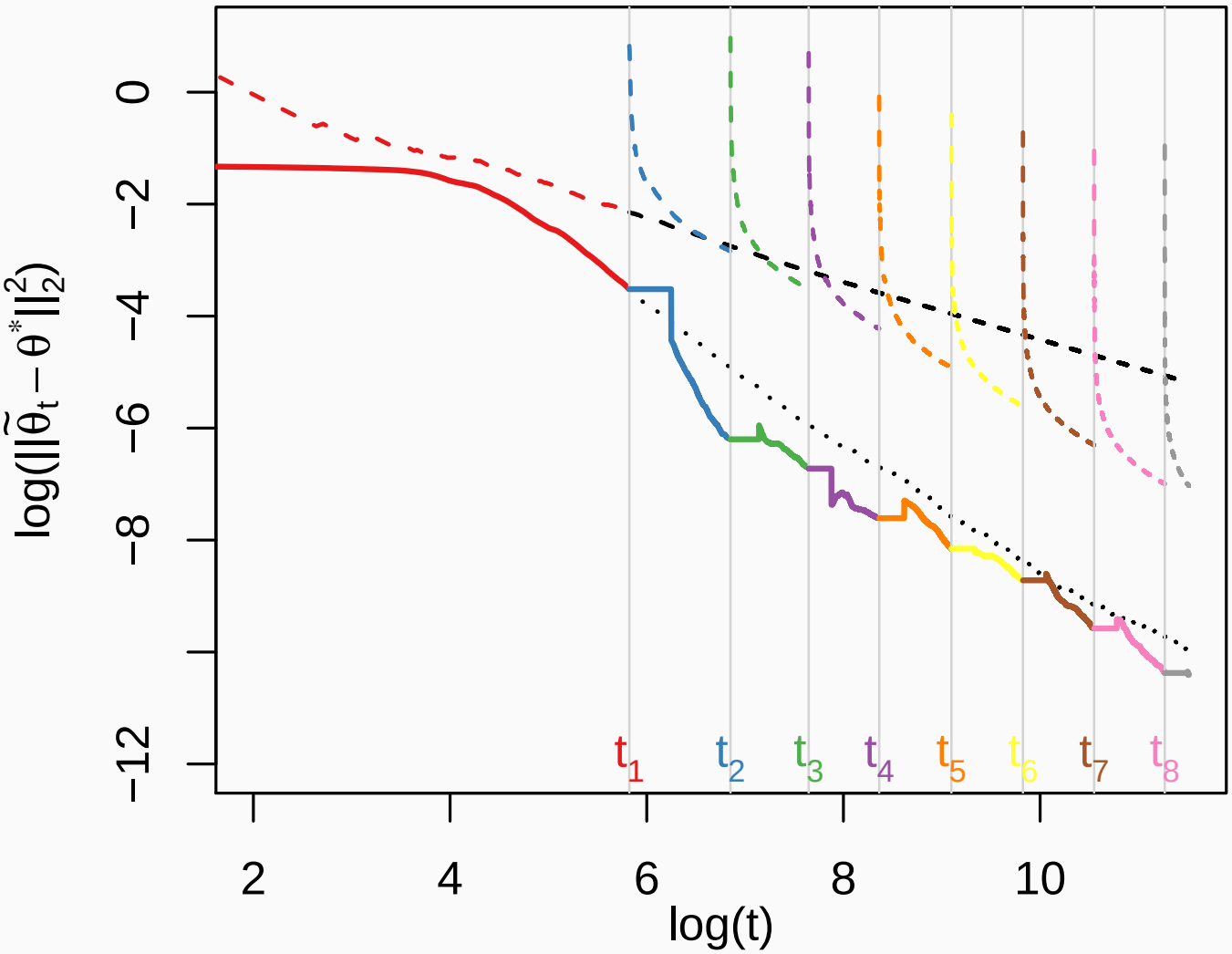
as soon as one θ_k is ε-close to θ^* in L1-norm.



Build an adaptive discretization

Enlarge the discretization set $\{\theta_1, \dots, \theta_K\}$ by sequentially adding the new estimators $\bar{\theta}_t$
$\rightarrow$ **auto-regulated procedure**

Acceleration in practice



Extensions: sparsity and Bernstein assumption

Sparsity: if θ^* is sparse with only d_0 non-zero coefficients,

$$Reg_n \lesssim \frac{d_0}{\mu} \frac{\log d}{n}$$

💡 add all truncation of $[\bar{\theta}_t]_s$ to its s largest component for $1 \leqslant s \leqslant d$ instead of just $\bar{\theta}_t$

Bernstein assumption: strong-convexity can be replaced with (weak assumptions and) the following assumption: it exists $0 \leqslant \beta \leqslant 1$ for all $\theta \in \mathrm{Support}(\theta^*)$

$$\mu \|\theta - \theta^*\|_2^2 \leqslant \mathbb{E}[\ell_t(\theta) - \ell_t(\theta^*)]^\beta$$

Then,

$$\mathrm{Reg}_n \lesssim \left(\frac{d_0 \log d}{\mu n} \right)^{\frac{1}{2-\beta}}$$

- 👍 continuity between strong-convexity ($\beta = 1$) and convexity $\beta = 0$
- 👍 the assumption only holds on $\mathrm{Support}(\theta^*) \Rightarrow \mu$ may be much larger
- 👍 the procedure is fully adaptive in d_0, β, μ

Work in progress and future work

- we assumed θ^* (approximately) sparse with $\|\theta^*\|_1 \leqslant U$ with a known bound $U > 0$
  Can we get an oracle bound of the form:

$$\widehat{L}_n \lesssim L_n(\theta) + \frac{\|\theta\|_0}{\mu} \frac{\log d}{n} \ ?$$

   Some solutions using projections.

- can we get rid of the i.i.d. assumption?

- can the algorithm be distributed?

THANK YOU !

References


Gaillard, P. and O. Wintenberger. “Sparse Accelerated Exponential Weights”. Accepted at AISTAT’17. 2017.


Littlestone, N. and M. K. Warmuth. “The weighted majority algorithm”. In: *Information and computation* 108.2 (1994), pp. 212–261.


Vovk, V. G. “Aggregating strategies”. In: *Proc. of Computational Learning Theory, 1990* (1990).