ACCÉLÉRATION PARCIMONIEUSE DES POIDS EXPONENTIELS

Pierre Gaillard

30 mai 2017

INRIA Paris – SIERRA

Travail en collaboration avec Olivier Wintenberger

CADRE

LE CADRE DE CET EXPOSÉ

Minimisation séquentielle d'une suite i.i.d. de fonctions de pertes[1] $\ell_1, \dots, \ell_n : \mathbb{R}^d \to \mathbb{R}$.

Cadre : À chaque pas de temps $t = 1, \dots, n$
- le statisticien met à jour son estimateur $\widehat{\theta}_{t-1} \in \mathbb{R}^d$ basé sur le passé $\ell_1, \dots, \ell_{t-1}$
- on observe ℓ_t.

Objectif : minimiser le risque de $\widehat{\theta}_n$

$$\mathrm{Risk}(\widehat{\theta}_n) := \mathbb{E}[\ell_{n+1}](\widehat{\theta}_n)\,.$$

Exemple : observation séquentielle d'un échantillon $\{(X_t, Y_t)\}_{1 \leqslant t \leqslant n}$ avec la perte carrée : $\ell_t : \theta \mapsto (Y_t - X_t \cdot \theta)^2$.

$$\mathrm{Risk}(\theta) = \mathbb{E}\left[(Y_t - X_t \cdot \theta)^2\right].$$

1. Cesa-Bianchi et Lugosi, *Prediction, Learning, and Games*, 2006.

Hypothèse de Bernstein (ou Lojasiewicz's)

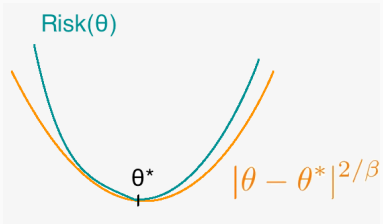
On suppose que le domaine, les gradients et pertes sont bornées ps. (ou sous-gaussien).

Hypothèse de Bernstein (ou Lojasiewicz's)

$$\exists \mu > 0, 0 \leqslant \beta \leqslant 1, \theta^* \in \mathbb{R}^d, \ \forall \theta \in \mathbb{R}^d \qquad \mu \left\| \theta - \theta^* \right\|_2^2 \leqslant \left[\text{Risk}(\theta) - \text{Risk}(\theta^*) \right]^{\beta}.$$

Remarques :

- Risque fortement convexe → vérifié avec $\beta = 1$
  → peut être vérifié pour des pertes non fortement convexes comme la perte absolue ou quantile, si les distributions sont assez régulières.
- Risque seulement convexe → vérifié avec $\mu = 1/(\text{Borne sur le domaine})$ et $\beta = 0$
- Tout un panel entre les deux même dans des cadres qui semblent discrets comme la classification.
  Ex : $Y \in \{-1, 1\}$, perte Hinge [a] → vérifié avec $\beta > 0$



a. Lecué, "Optimal Rates of Aggregation in Classification Under Low Noise Assumption", 2006.

Hypothèse de Bernstein (ou Lojasiewicz's)

On suppose que le domaine, les gradients et pertes sont bornées ps. (ou sous-gaussien).

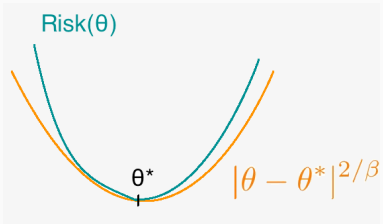
Hypothèse de Bernstein (ou Lojasiewicz's)

$$\exists \mu > 0, 0 \leqslant \beta \leqslant 1, \theta^* \in \mathbb{R}^d, \ \forall \theta \in \mathbb{R}^d \qquad \mu \left\| \theta - \theta^* \right\|_2^2 \leqslant \left[\text{Risk}(\theta) - \text{Risk}(\theta^*) \right]^\beta .$$

θ^* is d_0-sparse $\qquad$ $\|\theta^*\|_1 \leqslant U$

Remarques :

- Risque fortement convexe → vérifié avec $\beta = 1$
  → peut être vérifié pour des pertes non fortement convexes comme la perte absolue ou quantile, si les distributions sont assez régulières.
- Risque seulement convexe → vérifié avec $\mu = 1/(\text{Borne sur le domaine})$ et $\beta = 0$
- Tout un panel entre les deux même dans des cadres qui semblent discrets comme la classification.
  Ex : $Y \in \{-1, 1\}$, perte Hinge [a] → vérifié avec $\beta > 0$



a. Lecué, "Optimal Rates of Aggregation in Classification Under Low Noise Assumption", 2006.

Hypothèse de Bernstein (ou Lojasiewicz's)

On suppose que le domaine, les gradients et pertes sont bornées ps. (ou sous-gaussien).

Hypothèse de Bernstein (ou Lojasiewicz's)

$$\exists \mu > 0, 0 \leqslant \beta \leqslant 1, \theta^* \in \mathbb{R}^d, \ \forall \theta \in \mathbb{R}^d \qquad \mu \big\| \theta - \theta^* \big\|_2^2 \leqslant \big[\mathrm{Risk}(\theta) - \mathrm{Risk}(\theta^*) \big]^{\beta}.$$

θ^* is d_0-sparse $\qquad$ $\|\theta^*\|_1 \leqslant U$

Résultat : on propose une procédure qui redémarre des algorithmes de minimisation dans la boule ℓ_1 pour obtenir de meilleures vitesses (en n)

$$\mathrm{Risk}(\widehat{\theta}_n) \lesssim \mathrm{Risk}(\theta^*) + \min \left\{ \left(\frac{d_0 \log d}{\mu n} \right)^{\frac{1}{2-\beta}}, U \sqrt{\frac{\log d}{n}} \right\}.$$

Vitesse rapide : mieux pour n, μ grands
Vitesse lente : mieux pour n, μ petits

COMPARISON WITH THE LITTERATURE

Procedure	Sequentielle	Vitesse ($\beta = 1$)	Polynomiale	Hypothèse
Lasso [2]	✗	$\frac{d_0 \log d}{\mu n}$	✓	Bernstein [3]
EWA + sparsity patern [4]	✗	$\frac{d_0 \log d}{n}$	✗	Forte Conv.
SeqSEW [5]	✓	$\frac{d_0 \log d}{n}$	✗	Forte Conv.
ℓ_1-RDA method [6]	✓	$\frac{d}{n}$	✓	Forte Conv.
Agarwal [7], Steinhard [8]	✓	$\frac{d_0 \log d}{\mu n}$	✓	Forte Conv.
SAEW	✓	$\frac{d_0 \log d}{\mu n}$	✓	Bernstein

2. Bunea, Tsybakov et Wegkamp, “Aggregation for Gaussian regression”, 2007.
3. Pierre, Vincent et Guillaume, “Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions”, 2017.
4. Rigollet et Tsybakov, “Exponential screening and optimal rates of sparse estimation”, 2011.
5. Gerchinovitz, “Sparsity regret bounds for individual sequences in online linear regression”, 2013.
6. Xiao, “Dual averaging methods for regularized stochastic learning and online optimization”, 2010.
7. Agarwal, Negahban et Wainwright, “Stochastic optimization and sparse statistical recovery : Optimal algorithms for high dimensions”, 2012.
8. Steinhardt, Wager et Liang, “The Statistics of Streaming Sparse Regression”, 2014.

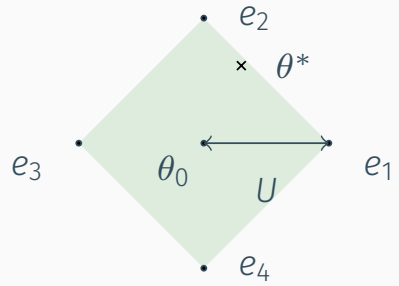
Optimisation convexe dans la boule ℓ_1-ball avec une vitesse lente

OPTIMISATION CONVEXE DANS LA BOULE ℓ_1-BALL AVEC UNE VITESSE LENTE

Obectif : approcher θ^* dans la boule ℓ_1 :

$$\mathcal{B}_1(\theta_0, U) := \{\theta \in \mathbb{R}^d : \|\theta - \theta_0\|_1 \leqslant U\}$$

On définie $e_k := \theta_0 + (0, \ldots, 0, \pm U, 0, \ldots, 0)$ pour $k = 1, \ldots, 2d$



L'algorithme *EG*$^{\pm}$ — Kivinen et Warmuth, 1997

Paramètres : tuning parameters $\eta_t > 0$.

pour *chaque itération* $t = 1, 2, \ldots$ **faire**

- Définir $\theta_{t-1} = \sum_{k=1}^{2d} p_{k,t} e_k$ et prévoir $\widehat{\theta}_{t-1} = \sum_{s=1}^{t-1} \theta_s$.
- Mettre à jour les poids pour $k = 1, \ldots, 2d$

$$p_{k,t} = \frac{e^{-\eta_t \sum_{s=1}^{t} \nabla \ell_s(\theta_{s-1}) \cdot e_k}}{\sum_{j=1}^{2d} e^{-\eta_t \sum_{s=1}^{t} \nabla \ell_s(\theta_{s-1}) \cdot e_j}}$$

fin

Performance : le choix $\eta_t \approx 1/(U\sqrt{t})$ assure avec grande probabilité

$$\mathrm{Risk}(\widehat{\theta}_n) - \mathrm{Risk}(\theta^*) \lesssim U\sqrt{\frac{\log d}{n}}\,.$$

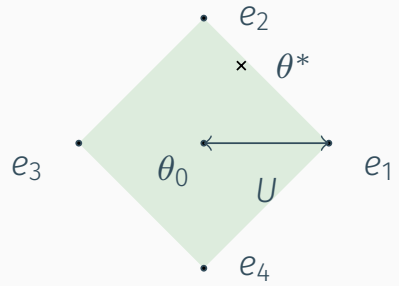
Remarque : beaucoup d'autres sous-routines peuvent être utilisées !

OPTIMISATION CONVEXE DANS LA BOULE ℓ_1-BALL AVEC UNE VITESSE LENTE

Obectif : approcher θ^* dans la boule ℓ_1 :

$$\mathcal{B}_1(\theta_0, U) := \{\theta \in \mathbb{R}^d : \|\theta - \theta_0\|_1 \leqslant U\}$$

On définie $e_k := \theta_0 + (0, \ldots, 0, \pm U, 0, \ldots, 0)$ pour $k = 1, \ldots, 2d$



L'algorithme *EG*$^{\pm}$ — Kivinen et Warmuth, 1997

Paramètres : tuning parameters $\eta_t > 0$.

pour *chaque itération* $t = 1, 2, \ldots$ **faire**

- Définir $\theta_{t-1} = \sum_{k=1}^{2d} p_{k,t} e_k$ et prévoir $\widehat{\theta}_{t-1} = \sum_{s=1}^{t-1} \theta_s$.
- Mettre à jour les poids pour $k = 1, \ldots, 2d$

$$p_{k,t} = \frac{e^{-\eta_t \sum_{s=1}^{t} \nabla \ell_s(\theta_{s-1}) \cdot e_k}}{\sum_{j=1}^{2d} e^{-\eta_t \sum_{s=1}^{t} \nabla \ell_s(\theta_{s-1}) \cdot e_j}}$$

fin

Performance : le choix $\eta_t \approx 1/(U\sqrt{t})$ assure avec grande probabilité

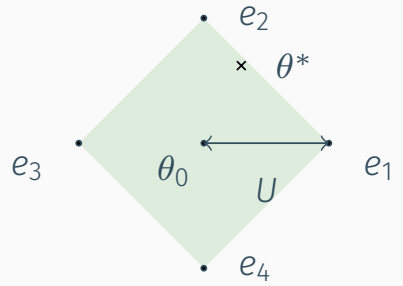
$$\mu \|\widehat{\theta}_n - \theta^*\|_2^{2/\beta} \leqslant \text{Risk}(\widehat{\theta}_n) - \text{Risk}(\theta^*) \lesssim U\sqrt{\frac{\log d}{n}}\,.$$

OPTIMISATION CONVEXE DANS LA BOULE ℓ_1-BALL AVEC UNE VITESSE LENTE

Obectif : approcher θ^* dans la boule ℓ_1 :

$$\mathcal{B}_1(\theta_0, U) := \{\theta \in \mathbb{R}^d : \|\theta - \theta_0\|_1 \leqslant U\}$$

On définie $e_k := \theta_0 + (0, \ldots, 0, \pm U, 0, \ldots, 0)$ pour $k = 1, \ldots, 2d$



L'algorithme *EG*$^{\pm}$ — Kivinen et Warmuth, 1997

Paramètres : tuning parameters $\eta_t > 0$.

pour *chaque itéraction* $t = 1, 2, \ldots$ **faire**

- Définir $\theta_{t-1} = \sum_{k=1}^{2d} p_{k,t} e_k$ et prévoir $\widehat{\theta}_{t-1} = \sum_{s=1}^{t-1} \theta_s$.
- Mettre à jour les poids pour $k = 1, \ldots, 2d$

$$p_{k,t} = \frac{e^{-\eta_t \sum_{s=1}^{t} \nabla \ell_s(\theta_{s-1}) \cdot e_k}}{\sum_{j=1}^{2d} e^{-\eta_t \sum_{s=1}^{t} \nabla \ell_s(\theta_{s-1}) \cdot e_j}}$$

fin

Performance : le choix $\eta_t \approx 1/(U\sqrt{t})$ assure avec grande probabilité

$$\left\|\widehat{\theta}_n - \theta^*\right\|_2 \lesssim \frac{1}{\sqrt{\mu}} \left(U\sqrt{\frac{\log d}{n}}\right)^{\frac{\beta}{2}} . \qquad (*)$$

Lemma (Hoeffding)

Si $-B \leqslant X \leqslant B$ est une variable aléatoire. Alors,

$$\forall \eta \in \mathbb{R}, \qquad \mathbb{E}[X] \leqslant -\frac{1}{\eta} \log\left(\mathbb{E}\left[e^{-\eta X}\right]\right) + \frac{\eta B}{4} .$$

1. On borne les gradients instantannés

$$\begin{aligned}
\nabla \ell_t(\theta_{t-1})^\top \theta_{t-1} &\overset{\text{def. of } \theta_{t-1}}{=} \sum_{k=1}^{2d} p_{k,t-1} \left(\nabla \ell_t(\theta_{t-1})^\top e_k\right) \\
&\overset{\text{Hoeffding}}{\leqslant} -\frac{1}{\eta} \log\left(\sum_{k=1}^{K} p_{k,t} e^{-\eta \nabla \ell_t(\theta_{t-1})^\top e_k}\right) + \frac{\eta B}{4} \\
&\overset{\text{def. of } p_{k,t+1}}{=} -\frac{1}{\eta} \log\left(\frac{p_{k,t}}{p_{k,t+1}} e^{-\eta \nabla \ell_t(\theta_{t-1})^\top e_k}\right) + \frac{\eta B}{4} \\
&= \nabla \ell_t(\theta_{t-1})^\top e_k + \frac{1}{\eta} \log \frac{p_{k,t+1}}{p_{k,t}} + \frac{\eta B}{4} .
\end{aligned}$$

2. On somme sur t, la somme télescope

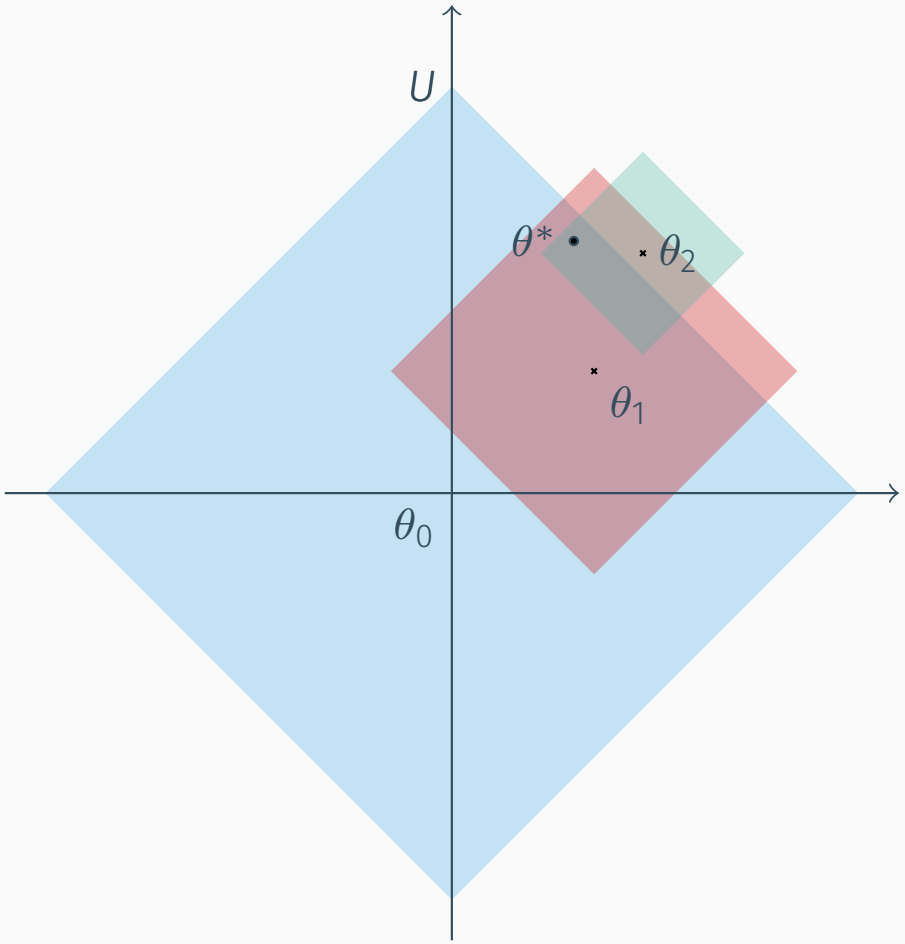
$$\begin{aligned}
\sum_{t=1}^{n} \ell_t(\theta_{t-1}) - \ell_t(\theta^*) &\overset{\text{Jensen}}{\leqslant} \sum_{1}^{n} \nabla \ell_t(\theta_{t-1})^\top (\theta_{t-1} - \theta^*) \leqslant \max_{k=1,\dots,2d} \left\{ \sum_{1}^{n} \nabla \ell_t(\theta_{t-1})^\top (\theta_{t-1} - e_k) \right\} \\
&\leqslant \max_{k=1,\dots,2d} \left\{ \frac{1}{\eta} \log \frac{p_{k,n}}{p_{k,0}} + \frac{\eta B n}{4} \right\} \leqslant \frac{\log(2d)}{\eta} + \frac{\eta B n}{4}
\end{aligned}$$

3. On majore le risque par le regret avec les inégalités d'Hoeffding et de Jensen

Accélération : de la vitesse lente à la vitesse rapide

ACCELERATION : DE LA VITESSE LENTE À LA VITESSSE RAPIDE

Idée : redémarrer régulièrement la sous-routine dans des boules de plus en plus petites autour de l'estimateur courant.



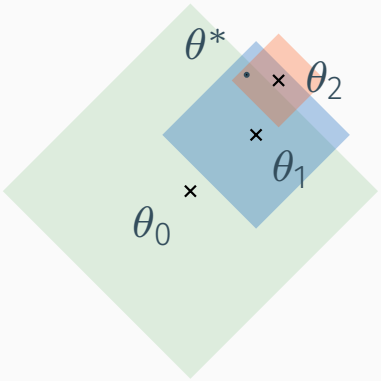
ACCELERATION : DE LA VITESSE LENTE À LA VITESSSE RAPIDE

Pour $i = 1, \dots, i_n$, **redémarrer** la sous-routine pour optimiser dans des boules de plus en plus petites de rayon $U_i := U2^{-i}$.

Où ? Centrées sur l'estimateur courant $\widehat{\theta}_t$.

Quand ? Quand on sait qu'avec grande probabilité

$$\left\|\widehat{\theta}_t - \theta^*\right\|_1 \leqslant U_i \,.$$



Calcul du quand (attention il faut un peu jongler) : d'après la vitesse lente :

$$\left\|\widehat{\theta}_{t_i} - \theta^*\right\|_1 \leqslant \sqrt{d}\left\|\widehat{\theta}_{t_i} - \theta^*\right\|_2 \overset{(*)}{\lesssim} \sqrt{\frac{d}{\mu}}\left(U_{i-1}\sqrt{\frac{\log d}{t_i}}\right)^{\frac{\beta}{2}} .$$

Comme $U_{i-1} = 2U_i$, il suffit donc que

$$\sqrt{\frac{d}{\mu}}\left(2U_i\sqrt{\frac{\log d}{t_i}}\right)^{\frac{\beta}{2}} \lesssim U_i \quad \Leftarrow \quad t_i \approx \left(\frac{d}{\mu U_i^{2-\beta}}\right)^{\frac{2}{\beta}} \log d$$

L'algorithme : redémarrer la sous-routine en $\widehat{\theta}_t$ tous les

$$t_i \approx \left(\frac{d}{\mu U_i^{2-\beta}}\right)^{\frac{2}{\beta}} \log d$$

instants.

Que se passe t-il ? La sous-routine travaille dans des boules de plus en plus petites,

$$U_i \approx \left(\frac{d \log d}{\mu t_i}\right)^{\frac{1}{2-\beta}} \sqrt{\frac{t_i}{\log d}}$$

L'algorithme : redémarrer la sous-routine en $\widehat{\theta}_t$ tous les

$$t_i \approx \left(\frac{d}{\mu U_i^{2-\beta}}\right)^{\frac{2}{\beta}} \log d$$

instants.

Que se passe t-il ? La sous-routine travaille dans des boules de plus en plus petites,

$$U_{i_n} \approx \left(\frac{d \log d}{\mu n}\right)^{\frac{1}{2-\beta}} \sqrt{\frac{n}{\log d}}$$

Transformation de la vitesse lente en vitesse rapide : la vitesse lente de la dernière sous-routine (au temps n) devient

$$\mathrm{Risk}(\widehat{\theta}_n) - \mathrm{Risk}(\theta^*) \lesssim U_{i_n} \sqrt{\frac{\log d}{n}} \lesssim \left(\frac{d \log d}{\mu n}\right)^{\frac{1}{2-\beta}}$$

L'algorithme : redémarrer la sous-routine en $\widehat{\theta}_t$ tous les

$$t_i \approx \left(\frac{d}{\mu U_i^{2-\beta}}\right)^{\frac{2}{\beta}} \log d$$

instants.

Que se passe t-il ? La sous-routine travaille dans des boules de plus en plus petites,

$$U_{i_n} \approx \left(\frac{d \log d}{\mu n}\right)^{\frac{1}{2-\beta}} \sqrt{\frac{n}{\log d}}$$

Transformation de la vitesse lente en vitesse rapide : la vitesse lente de la dernière sous-routine (au temps n) devient

$$\mathrm{Risk}(\widehat{\theta}_n) - \mathrm{Risk}(\theta^*) \lesssim U_{i_n} \sqrt{\frac{\log d}{n}} \lesssim \left(\frac{d \log d}{\mu n}\right)^{\frac{1}{2-\beta}}$$

Mais on ne voulait pas d_0 plutôt que d ?

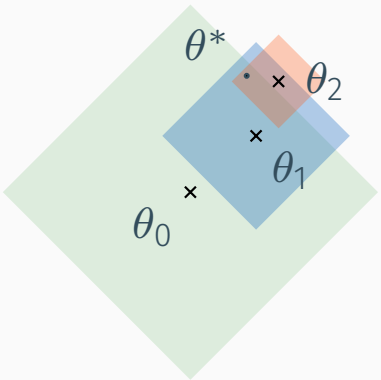
ACCELERATION : DE LA VITESSE LENTE À LA VITESSSE RAPIDE

Pour $i = 1, \dots, i_n$, **redémarrer** la sous-routine pour optimiser dans des boules de plus en plus petites de rayon $U_i := U2^{-i}$.

Où ? Centrées sur l'estimateur courant tronqué $[\widehat{\theta}_t]_{d_0}$ à d_0 coordonnées non-nulles.

Quand ? Quand on sait qu'avec grande probabilité

$$\left\| [\widehat{\theta}_t]_{d_0} - \theta^* \right\|_1 \leqslant U_i \, .$$



Calcul du quand (attention il faut un peu jongler) : d'après la vitesse lente :

$$\left\| \widehat{\theta}_{t_i} - \theta^* \right\|_1 \leqslant \sqrt{d_0} \left\| \widehat{\theta}_{t_i} - \theta^* \right\|_2 \overset{(*)}{\lesssim} \sqrt{\frac{d_0}{\mu}} \left(U_{i-1} \sqrt{\frac{\log d}{t_i}} \right)^{\frac{\beta}{2}} .$$

Les même calculs [...] donnent

$$\mathrm{Risk}(\widehat{\theta}_n) - \mathrm{Risk}(\theta^*) \lesssim U_{i_n} \sqrt{\frac{\log d}{n}} \lesssim \left(\frac{d_0 \log d}{\mu n} \right)^{\frac{1}{2-\beta}}$$

VITESSE LENTE À VITESSE RAPIDE : ILLUSTRATION

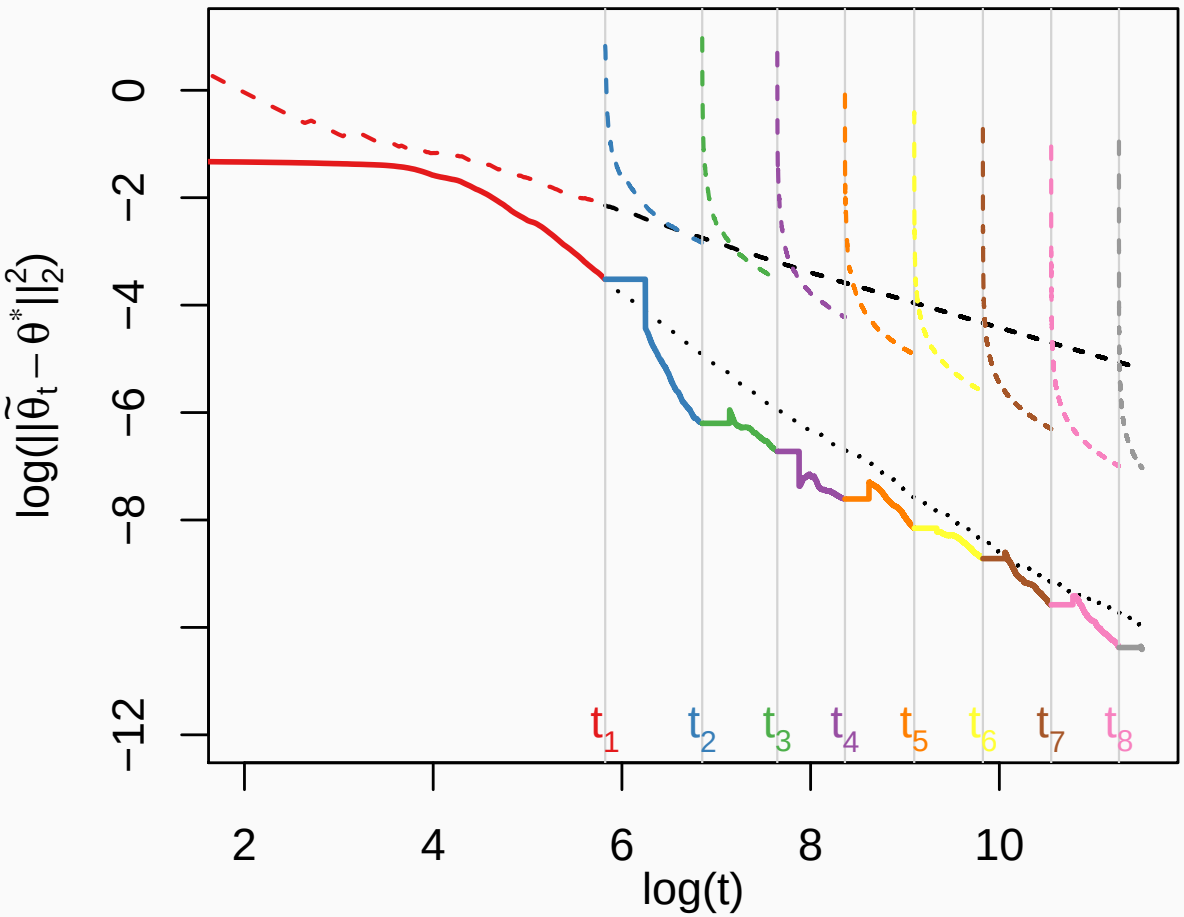


FIGURE 1 – Logarithm of the ℓ_2-error of the averaged estimator.

SIMULATIONS : RÉGRESSION LINÉAIRE DES MOINDRES CARRÉS

Échantillon $(X_t, Y_t) \in [-X, X]^d \times [-Y, Y]$ i.i.d.

Objectif : estimer linéairement Y_t en approchant

$$\theta^* \in \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \mathbb{E}\left[(Y_t - X_t^\top \theta)^2\right]$$

Le risque est fortement convexe $\beta = 1$ avec $\mu \leqslant \lambda_{\min}\left(\mathbb{E}\left[X_t X_t^\top\right]\right)$.

Expérience : $X_t \sim \mathcal{N}(0, 1)$ for $d = 500$, $n = 2\,000$

$$Y_t = X_t^\top \theta^* + 0.1\,\varepsilon_t \qquad \text{avec} \quad \varepsilon_t \sim \mathcal{N}(0, 1) \quad \text{i.i.d.}$$

où $d_0 = \|\theta^*\|_0 = 5$, $U = \|\theta^*\|_1 = 1$, coordonnées non nulles i.i.d. $\propto \mathcal{N}(0, 1)$, $\mu = 1$.

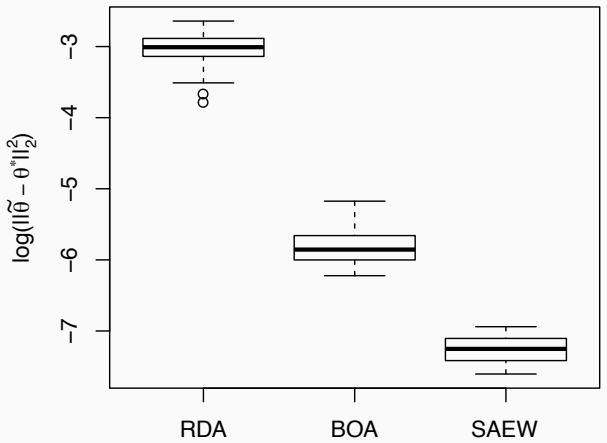


FIGURE 2 – Boxplot (30 simulations)

Simulations : $d_0 = 5$, $d = 500$, $\sigma = 0.1$

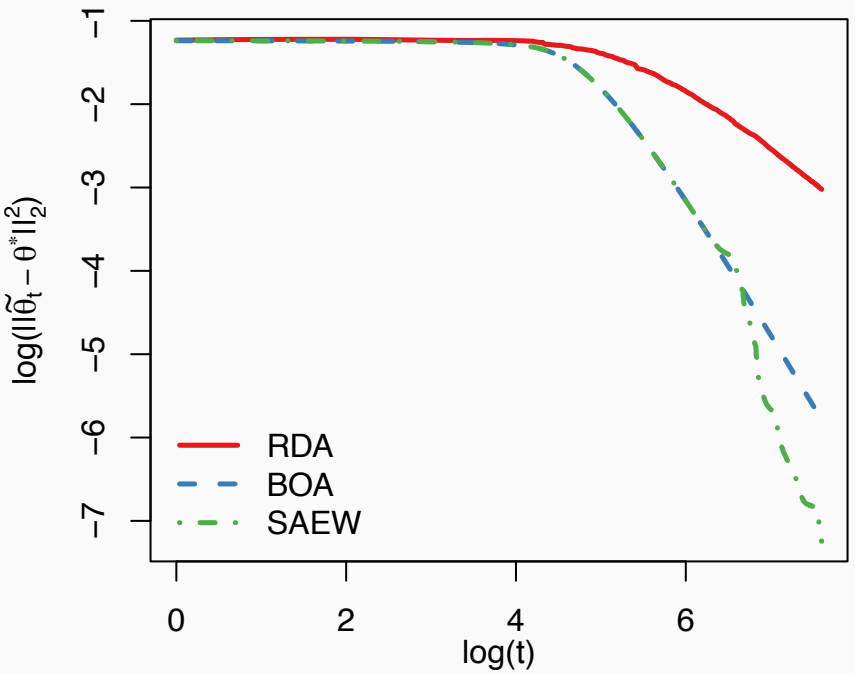


FIGURE 3 – Log de l'erreur ℓ_2.

En pratique : bien meilleur si on autorise plusieurs passes sur les données.

Travail en court et futur :

- Calibration des paramètres plus propre (sans surcouche) : procédure complètement adaptative.
- Borne oracle : pas d'hypothèse de parcimonie sur θ^*
- Le faire sans recommencer ?
- Applications...

Merci !

RÉFÉRENCES


Agarwal, A., S. Negahban et M. J. Wainwright. “Stochastic optimization and sparse statistical recovery : Optimal algorithms for high dimensions”. In : *Advances in Neural Information Processing Systems*. 2012, p. 1538–1546.


Bunea, F., A. Tsybakov et M. Wegkamp. “Aggregation for Gaussian regression”. In : *The Annals of Statistics* 35.4 (2007), p. 1674–1697.


Cesa-Bianchi, N. et G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.


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


Gerchinovitz, S. “Sparsity regret bounds for individual sequences in online linear regression”. In : *The Journal of Machine Learning Research* 14.1 (2013), p. 729–769.


Lecué, G. “Optimal Rates of Aggregation in Classification Under Low Noise Assumption”. Thèse de doct. 2006.


Pierre, A., C. Vincent et L. Guillaume. “Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions”. In : *arXiv preprint arXiv :1702.01402* (2017).


Rigollet, P. et A. Tsybakov. “Exponential screening and optimal rates of sparse estimation”. In : *The Annals of Statistics* (2011), p. 731–771.


Steinhardt, J., S. Wager et P. Liang. “The Statistics of Streaming Sparse Regression”. In : *arXiv preprint arXiv :1412.4182* (2014).


Xiao, L. “Dual averaging methods for regularized stochastic learning and online optimization”. In : *Journal of Machine Learning Research* 11 (2010), p. 2543–2596.