Regularization with Lipschitz Loss

Pierre Alquier

École nationale de la statistique et de l'administration économique
ENSAE ParisTech
université PARIS-SACLAY

Sequential, structured, and/or statistical learning
IHES - May 17, 2017

Motivation : user ratings

									
Stan			7		3		8		
Pierre	8	10	9	10	9	10	10	10	8
Zoe	8	3					7		
Bob			6	4				2	
Oscar				6		10		7	
Léa		8	4		9				
Tony			9	3				4	8

Motivation : user ratings

									
Stan			7		3		8		
Pierre	8	10	9	10	9	10	10	10	8
Zoe	8	3					7		
Bob			6	4				2	? ? ?
Oscar				6		10		7	
Léa		8	4		9				
Tony			9	3				4	8

Motivation : user ratings

									
Stan			7		3		8		
Pierre	8	10	9	10	9	10	10	10	8
Zoe	8	3					7		
Bob			6	4				2	7
Oscar				6		10		7	
Léa		8	4		9				
Tony			9	3				4	8

A possible model

Notation : $\langle A, B \rangle_F = \mathrm{Tr}(A^T B)$. Let $E_{j,k}$ be the matrix with zeros everywhere except the (j,k)-th entry equal to 1.

A possible model

Notation : $\langle A, B \rangle_F = \mathrm{Tr}(A^T B)$. Let $E_{j,k}$ be the matrix with zeros everywhere except the (j,k)-th entry equal to 1.

Observations :

$$Y_i = \langle M^*, X_i \rangle_F + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i) = 0$$

X_i takes values in the set of matrices $\{E_{j,k}\}$.

A possible model

Notation : $\langle A, B \rangle_F = \mathrm{Tr}(A^T B)$. Let $E_{j,k}$ be the matrix with zeros everywhere except the (j,k)-th entry equal to 1.

Observations :

$$Y_i = \langle M^*, X_i \rangle_F + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i) = 0$$

X_i takes values in the set of matrices $\{E_{j,k}\}$.

Idea : M^* is (approximately) low-rank.

E. Candès & T. Tao (2009). The power of convex relaxation : Near-optimal matrix completion. *IEEE Trans. Info. Theory.*

E. Candès & Y. Plan (2010). Matrix completion with noise. *Proceedings of the IEEE.*

Penalized ERM

First idea :

$$\hat{M} \in \arg\min \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \langle M, X_i \rangle_F \right)^2 + \lambda . \mathrm{rank}(M) \right\}$$

but the rank is not convex...

Penalized ERM

First idea :

$$\hat{M} \in \arg\min \left\{ \frac{1}{N} \sum_{i=1}^{N} (Y_i - \langle M, X_i \rangle_F)^2 + \lambda . \mathrm{rank}(M) \right\}$$

but the rank is not convex...

$$\hat{M} \in \arg\min \left\{ \frac{1}{N} \sum_{i=1}^{N} (Y_i - \langle M, X_i \rangle_F)^2 + \lambda \|M\|_* \right\}$$

Minimax rates of convergence derived in

V. Koltchinskii, K. Lounici, & A. Tsybakov (2011) Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*.

O. Klopp (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*.

Is the quadratic loss always a good idea ?

									
Stan			7		3		8		
Pierre	8	10	9	10	9	10	10	10	8
Zoe	8	3					7		
Bob			6	4				2	
Oscar				6		10		7	
Léa		8	4		9				
Tony			9	3				4	8

Is the quadratic loss always a good idea ?

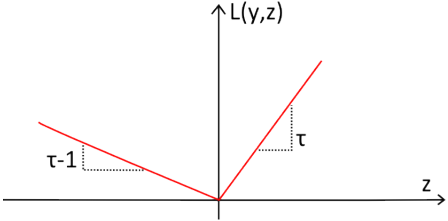
									
Stan			7		3		8		
Pierre	8	10	9	10	9	10	10	10	8
Zoe	8	3					7		
Bob			6	4				2	???
Oscar				6		10		7	
Léa		8	4		9				
Tony			9	3				4	8

Is the quadratic loss always a good idea ?

									
Stan			7		3		8		
Pierre	8	10	9	10	9	10	10	10	8
Zoe	8	3					7		
Bob			6	4				2	[6,8]
Oscar				6		10		7	
Léa		8	4		9				
Tony			9	3				4	8

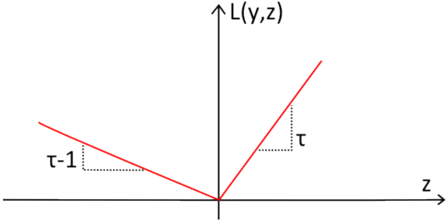
The quantile loss

... suggests to replace the quadratic loss by the quantile loss $\ell_\tau(f(x), y) = (y - f(x))[\tau - \mathbf{1}(y - f(x) \leq 0)]$.



The quantile loss

... suggests to replace the quadratic loss by the quantile loss $\ell_\tau(f(x), y) = (y - f(x))[\tau - \mathbf{1}(y - f(x) \leq 0)]$.



$$\hat{M} \in \arg\min \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell_\tau(\langle M, X_i \rangle_F , Y_i) + \lambda \|M\|_* \right\}$$

Source : http ://www.lokad.com/

1-bit matrix completion

									
Stan			👍		👎		👍		
Pierre		👍	👍	👍	👍	👍	👍	👍	👍
Zoe	👍	👎					👍		
Bob			👍	👎				👎	
Oscar				👍		👍		👍	
Léa		👍	👎		👍				
Tony			👍	👎				👎	👍

1-bit matrix completion

									
Stan			👍		👎		👍		
Pierre		👍	👍	👍	👍	👍	👍	👍	👍
Zoe	👍	👎					👍		
Bob			👍	👎				👎	? ? ?
Oscar				👍		👍		👍	
Léa		👍	👎		👍				
Tony			👍	👎				👎	👍

1-bit matrix completion

									
Stan			👍		👎		👍		
Pierre		👍	👍	👍	👍	👍	👍	👍	👍
Zoe	👍	👎					👍		
Bob			👍	👎				👎	👍
Oscar				👍		👍		👍	
Léa		👍	👎		👍				
Tony			👍	👎				👎	👍

1-bit matrix completion

$$\hat{M} \in \arg\min \left\{ \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\mathrm{sign}(\langle M, X_i \rangle_F) \neq Y_i) + \lambda \|M\|_* \right\}$$

1-bit matrix completion

$$\hat{M} \in \arg\min \left\{ \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\mathrm{sign}(\langle M, X_i \rangle_F) \neq Y_i) + \lambda \|M\|_* \right\}$$

Problem : the indicator function is not convex.

1-bit matrix completion

$$\hat{M} \in \arg\min \left\{ \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\mathrm{sign}(\langle M, X_i \rangle_F) \neq Y_i) + \lambda \|M\|_* \right\}$$

Problem : the indicator function is not convex.

$$\hat{M} \in \arg\min \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell(\langle M, X_i \rangle_F \, , Y_i) + \lambda \|M\|_* \right\}$$

- logistic loss $\ell(y', y) = \log(1 + \exp(-y'y))$

J. Laffond, O. Klopp, E. Moulines & J. Salmon (2014). Probabilistic low-rank matrix completion on finite alphabets. *NIPS*.

1-bit matrix completion

$$\hat{M} \in \arg\min \left\{ \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\mathrm{sign}(\langle M, X_i \rangle_F) \neq Y_i) + \lambda \|M\|_* \right\}$$

Problem : the indicator function is not convex.

$$\hat{M} \in \arg\min \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell(\langle M, X_i \rangle_F , Y_i) + \lambda \|M\|_* \right\}$$

- logistic loss $\ell(y', y) = \log(1 + \exp(-y'y))$

  J. Laffond, O. Klopp, E. Moulines & J. Salmon (2014). Probabilistic low-rank matrix completion on finite alphabets. *NIPS*.

- hinge loss $\ell(y', y) = (1 - y'y)_+$ etc.

Lipschitz losses

All the aforementionned losses :

- hinge,
- logistic,
- quantile

are Lipschitz. And so are other popular losses :

- Huber,
- ...

Outline of the talk

1. Motivation
   - Matrix completion : the L_2 point of view
   - Matrix completion : Lipschitz losses ?
2. Oracle inequalities
   - Notations and overview
   - The main ingredients
   - Sharp oracle inequality
3. Applications
   - Logistic LASSO
   - Logistic SLOPE
   - Matrix completion with hinge loss

Outline of the talk

1. Motivation
   - Matrix completion : the L_2 point of view
   - Matrix completion : Lipschitz losses ?
2. Oracle inequalities
   - Notations and overview
   - The main ingredients
   - Sharp oracle inequality
3. Applications
   - Logistic LASSO
   - Logistic SLOPE
   - Matrix completion with hinge loss

Notations

- Pairs $(X_1, Y_1), \ldots, (X_N, Y_N)$ in $\mathcal{X} \times \mathbb{R}$ i.i.d from P.

Notations

- Pairs $(X_1, Y_1), \dots, (X_N, Y_N)$ in $\mathcal{X} \times \mathbb{R}$ i.i.d from P.
- A space $E \subseteq L_2(P)$ of functions $f : \mathcal{X} \to \mathbb{R}$ equipped with a norm $\| \cdot \|$, generally different from $\| \cdot \|_{L_2}$. A convex $F \subseteq E$.

Notations

- Pairs $(X_1, Y_1), \ldots, (X_N, Y_N)$ in $\mathcal{X} \times \mathbb{R}$ i.i.d from P.
- $F \subseteq E \subseteq L_2(P)$, $(E, \|\cdot\|)$.
- A loss function ℓ that is 1-Lipschitz :

$$|\ell(f_1(x), y) - \ell(f_2(x), y)| \leq |f_1(x) - f_2(x)|.$$

Notations

- Pairs $(X_1, Y_1), \ldots, (X_N, Y_N)$ in $\mathcal{X} \times \mathbb{R}$ i.i.d from P.
- $F \subseteq E \subseteq L_2(P)$, $(E, \| \cdot \|)$.
- A loss function ℓ
- oracle

$$f^* \in \arg\min_{f \in F} \underbrace{\mathbb{E}_P[\ell(f(X), Y)]}_{=R(f)}.$$

Notations

- Pairs $(X_1, Y_1), \ldots, (X_N, Y_N)$ in $\mathcal{X} \times \mathbb{R}$ i.i.d from P.
- $F \subseteq E \subseteq L_2(P)$, $(E, \|\cdot\|)$.
- A loss function ℓ
- oracle $f^* \in \arg\min_{f \in F} R(f)$.
- estimator :

Penalized ERM

$$\hat{f} \in \arg\min_{f \in F} \left[\frac{1}{N} \sum_{i=1}^{N} \ell(f(X_i), Y_i) + \lambda \|f\| \right].$$

Three main ingredients to study $\hat{f}$

- The Bernstein condition with parameters A and κ quantifies the "identifiability" of f^*.

Three main ingredients to study $\hat{f}$

- The Bernstein condition with parameters A and κ .
- The complexity parameter $\mathrm{comp}(B)$ measures the "size" or "complexity" of (the unit ball B of) E. Allows to define the complexity function

$$r(\rho) = \left[\frac{\rho A \mathrm{comp}(B)}{\sqrt{N}}\right]^{\frac{1}{2\kappa}} .$$

Three main ingredients to study $\hat{f}$

- The Bernstein condition with parameters A and κ .
- The complexity function

$$r(\rho) = \left[\frac{\rho A \mathrm{comp}(B)}{\sqrt{N}}\right]^{\frac{1}{2\kappa}} .$$

- The sparsity function $\Delta(\cdot)$ measures the size of the sub-differential of $\|\cdot\|$ in a ρ-neighborhood of f^*. Find a solution ρ^* to the sparsity equation

$$\Delta(\rho^*) \geq (4/5)\rho^*.$$

Three main ingredients to study $\hat{f}$

- The Bernstein condition with parameters A and κ .
- The complexity function

$$r(\rho) = \left[\frac{\rho A \mathrm{comp}(B)}{\sqrt{N}}\right]^{\frac{1}{2\kappa}} .$$

- Find a solution ρ^* to the sparsity equation

$$\Delta(\rho^*) \geq (4/5)\rho^*.$$

Then with high probability,

$$\|\hat{f} - f^*\| \leq \rho^*, \; \|\hat{f} - f^*\|_{L_2} \leq r(2\rho^*),$$

$$R(\hat{f}) - R(f^*) \lesssim [r(2\rho^*)]^{2\kappa}.$$

The Bernstein condition

The Bernstein condition

There is $\kappa \geq 1$ and $A > 0$ such that

$$\forall f \in F, \quad \|f - f^*\|_{L_2}^{2\kappa} \leq A[R(f) - R(f^*)].$$

The Bernstein condition and strongly convex losses

$$\forall f \in F, \quad \|f - f^*\|_{L_2}^{2\kappa} \leq A[R(f) - R(f^*)].$$

The Bernstein condition and strongly convex losses

$$\forall f \in F, \quad \|f - f^*\|_{L_2}^{2\kappa} \leq A[R(f) - R(f^*)].$$

P. Bartlett, M. Jordan & J. McAuliffe (2006). Convexity, classification and risk bounds. *JASA*.

Theorem

ℓ is strongly convex $\Rightarrow$ condition satisfied with $\kappa = 1$.

The Bernstein condition and strongly convex losses

$$\forall f \in F, \quad \|f - f^*\|_{L_2}^{2\kappa} \leq A[R(f) - R(f^*)].$$

P. Bartlett, M. Jordan & J. McAuliffe (2006). Convexity, classification and risk bounds. *JASA*.

Theorem

ℓ is strongly convex $\Rightarrow$ condition satisfied with $\kappa = 1$.

Proof :

The Bernstein condition and strongly convex losses

$$\forall f \in F, \quad \|f - f^*\|_{L_2}^{2\kappa} \leq A[R(f) - R(f^*)].$$

P. Bartlett, M. Jordan & J. McAuliffe (2006). Convexity, classification and risk bounds. *JASA*.

Theorem

ℓ is strongly convex $\Rightarrow$ condition satisfied with $\kappa = 1$.

Proof :

$$\frac{\ell(f(X),Y) + \ell(f^*(X),Y)}{2} - \ell\left(\frac{f(X)+f^*(X)}{2}, Y\right) \geq \alpha[f(X) - f^*(X)]^2.$$

The Bernstein condition and strongly convex losses

$$\forall f \in F, \quad \|f - f^*\|_{L_2}^{2\kappa} \leq A[R(f) - R(f^*)].$$

P. Bartlett, M. Jordan & J. McAuliffe (2006). Convexity, classification and risk bounds. *JASA*.

Theorem

ℓ is strongly convex $\Rightarrow$ condition satisfied with $\kappa = 1$.

Proof :

$$\frac{\ell(f(X), Y) + \ell(f^*(X), Y)}{2} - \ell\left(\frac{f(X) + f^*(X)}{2}, Y\right) \geq \alpha[f(X) - f^*(X)]^2.$$

$$\frac{R(f) + R(f^*)}{2} - \underbrace{R\left(\frac{f + f^*}{2}\right)} \geq \alpha\|f - f^*\|_{L_2}^2.$$

The Bernstein condition and strongly convex losses

$$\forall f \in F, \quad \|f - f^*\|_{L_2}^{2\kappa} \leq A[R(f) - R(f^*)].$$

P. Bartlett, M. Jordan & J. McAuliffe (2006). Convexity, classification and risk bounds. *JASA*.

Theorem

ℓ is strongly convex $\Rightarrow$ condition satisfied with $\kappa = 1$.

Proof :

$$\frac{\ell(f(X),Y) + \ell(f^*(X),Y)}{2} - \ell\left(\frac{f(X)+f^*(X)}{2}, Y\right) \geq \alpha[f(X) - f^*(X)]^2.$$

$$\frac{R(f)+R(f^*)}{2} - \underbrace{R\left(\frac{f+f^*}{2}\right)}_{\geq R(f^*)} \geq \alpha\|f - f^*\|_{L_2}^2.$$

The Bernstein condition and strongly convex losses

$$\forall f \in F, \quad \|f - f^*\|_{L_2}^{2\kappa} \leq A[R(f) - R(f^*)].$$

P. Bartlett, M. Jordan & J. McAuliffe (2006). Convexity, classification and risk bounds. *JASA*.

Theorem

ℓ is strongly convex $\Rightarrow$ condition satisfied with $\kappa = 1$.

Proof :

$$\frac{\ell(f(X), Y) + \ell(f^*(X), Y)}{2} - \ell\left(\frac{f(X) + f^*(X)}{2}, Y\right) \geq \alpha[f(X) - f^*(X)]^2.$$

$$\frac{R(f) + R(f^*)}{2} - \underbrace{R\left(\frac{f + f^*}{2}\right)}_{\geq R(f^*)} \geq \alpha\|f - f^*\|_{L_2}^2.$$

$$R(f) - R(f^*) \geq 2\alpha\|f - f^*\|_{L_2}^2.$$

The Bernstein condition and the hinge loss

$$\forall f \in F, \quad \|f - f^*\|_{L_2}^{2\kappa} \leq A[R(f) - R(f^*)].$$

The Bernstein condition and the hinge loss

$$\forall f \in F, \quad \|f - f^*\|_{L_2}^{2\kappa} \leq A[R(f) - R(f^*)].$$

G. Lecué (2006). Optimal Rates of Aggregation in Classification Under Low Noise Assumption. *PhD Thesis*.

Theorem

$Y \in \{-1, 1\}$, $\eta(x) := \mathbb{E}(Y|X = x)$ and $f^*(x) = \mathrm{sign}(\eta(x))$.

- $|\eta(X)| \geq \tau > 0$ a.s. $\Rightarrow$ Bernstein condition with $\kappa \geq 1$.

The Bernstein condition and the hinge loss

$$\forall f \in F, \quad \|f - f^*\|_{L_2}^{2\kappa} \leq A[R(f) - R(f^*)].$$

G. Lecué (2006). Optimal Rates of Aggregation in Classification Under Low Noise Assumption. *PhD Thesis*.

Theorem

$Y \in \{-1, 1\}$, $\eta(x) := \mathbb{E}(Y|X = x)$ and $f^*(x) = \mathrm{sign}(\eta(x))$.

- $|\eta(X)| \geq \tau > 0$ a.s. $\Rightarrow$ Bernstein condition with $\kappa \geq 1$.
- $\mathbb{P}(|\eta(X)| \leq t) \leq ct^{\frac{1}{\kappa-1}}$ with $\kappa > 1 \Rightarrow$ Bernstein.

The complexity parameter 1 - the bounded case

Let us assume that $\sup_{f\in F} \|f\|_\infty \leq b$.

The complexity parameter 1 - the bounded case

Let us assume that $\sup_{f\in F}\|f\|_\infty \leq b$.

Ex : matrix completion case, $X_i \in \{E_{j,k}\}$ and

$$F = \left\{ \langle M, \cdot \rangle_F , \sup_{i,j} |M_{i,j}| \leq b \right\}.$$

The complexity parameter 1 - the bounded case

Let us assume that $\sup_{f\in F}\|f\|_{\infty}\leq b$.

Ex : matrix completion case, $X_i \in \{E_{j,k}\}$ and

$$F = \left\{ \langle M, \cdot \rangle_F , \sup_{i,j} |M_{i,j}| \leq b \right\}.$$

Rademacher complexity

In this case we define, for B the unit ball in E,

$$\mathrm{comp}(B) = \mathbb{E} \sup_{f\in B} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_i f(X_i) \right|, \; (\epsilon_i) \text{ i.i.d Rademacher.}$$

The complexity parameter 2 - subgaussian case

Put $H = \{f - g, (f,g) \in F^2\}$. Assume that $\forall h \in H$, $\forall \lambda$,

$$\mathbb{E} \exp\left(\lambda \frac{|h(X)|}{\|h\|_{L_2}} \right) \leq \exp\left(\lambda^2 L^2\right).$$

The complexity parameter 2 - subgaussian case

Put $H = \{f - g, (f,g) \in F^2\}$. Assume that $\forall h \in H$, $\forall \lambda$,

$$\mathbb{E} \exp\left(\lambda \frac{|h(X)|}{\|h\|_{L_2}}\right) \leq \exp\left(\lambda^2 L^2\right).$$

Ex : X is Gaussian and

$$F = \{\langle t, \cdot \rangle, t \in \mathbb{R}^p\}.$$

The complexity parameter 2 - subgaussian case

Put $H = \{f - g, (f,g) \in F^2\}$. Assume that $\forall h \in H$, $\forall \lambda$,

$$\mathbb{E} \exp\left(\lambda \frac{|h(X)|}{\|h\|_{L_2}}\right) \leq \exp\left(\lambda^2 L^2\right).$$

Ex : X is Gaussian and

$$F = \left\{\langle t, \cdot \rangle, t \in \mathbb{R}^p\right\}.$$

Gaussian mean width

$(G_h)_{h \in E}$ canonical Gaussian process,

$$\mathrm{comp}(B) = \mathbb{E} \sup_{h \in B} G_h.$$

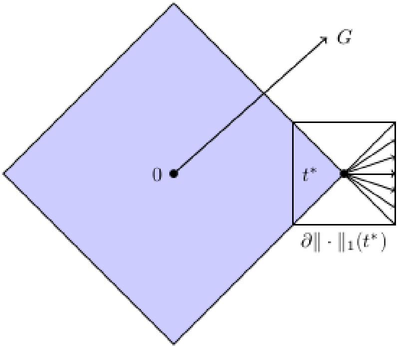
The complexity function

The complexity function

$$r(\rho) := \left[\frac{A\rho \mathrm{comp}(B)}{\sqrt{N}} \right]^{\frac{1}{2\kappa}} .$$

The sparsity equation

Example : the $\|\cdot\|_1$ penalty.



Idea : t^* sparse (easier to estimate) $\leftrightarrow$ $\partial\|\cdot\|(t^*)$ is a large set.

The sparsity equation

The sparsity parameter

$$\Delta(\rho) := \inf_{h \in \rho S \cap r(2\rho) B_{L_2}} \sup_{f \in \partial \|\cdot\|(f^*)} \langle h, f \rangle$$

where B_{L_2} is the unit ball in L_2 and S is the unit sphere in E.

The sparsity equation

The sparsity parameter

$$\Delta(\rho) := \inf_{h \in \rho S \cap r(2\rho) B_{L_2}} \sup_{f \in \partial \|\cdot\|(f^*)} \langle h, f \rangle$$

where B_{L_2} is the unit ball in L_2 and S is the unit sphere in E.

The sparsity equation

Find (the smallest possible) ρ^* such that

$$\Delta(\rho^*) \geq (4/5)\rho^*$$

Sharp oracle inequality

C stands for a constant that depends on A, κ, ... and may change from line to line.

Theorem

Take $\lambda = 720\mathrm{comp}(B)/(7\sqrt{N})$. Then with probability at least

$$1 - C\exp\left[-CN^{\frac{1}{2\kappa}}\left(\rho^{*}\mathrm{comp}(B)\right)^{\frac{2\kappa-1}{\kappa}}\right]$$

we have simultaneously

$$\|\hat{f} - f^{*}\| \leq \rho^{*},\ \|\hat{f} - f^{*}\|_{L_2} \leq r(2\rho^{*}),$$

$$R(\hat{f}) - R(f^{*}) \leq C[r(2\rho^{*})]^{2\kappa}.$$

Outline of the talk

Outline of the talk

1. Motivation
   - Matrix completion : the L_2 point of view
   - Matrix completion : Lipschitz losses ?
2. Oracle inequalities
   - Notations and overview
   - The main ingredients
   - Sharp oracle inequality
3. Applications
   - Logistic LASSO
   - Logistic SLOPE
   - Matrix completion with hinge loss

Logistic LASSO : context

$E = F = \{\langle t, \cdot \rangle, t \in \mathbb{R}^p\}$ equipped with $\| \cdot \| = \| \cdot \|_1$.

Logistic LASSO

$$\hat{f} \in \arg\min_{f \in F} \left[\frac{1}{N} \sum_{i=1}^{N} \log(1 - \exp(-Y_i f(X_i))) + \lambda \|f\|_1 \right].$$

Logistic LASSO : Bernstein & complexity

Assume that $X \sim \mathcal{N}(0, I_p)$.

Logistic LASSO : Bernstein & complexity

Assume that $X \sim \mathcal{N}(0, I_p)$.

Bernstein condition satisfied with $\kappa = 1$.

Logistic LASSO : Bernstein & complexity

Assume that $X \sim \mathcal{N}(0, I_p)$.

Bernstein condition satisfied with $\kappa = 1$.

$$\mathrm{comp}(B) = \mathbb{E} \sup_{\|t\|_1 \leq 1} \langle t, X \rangle = \mathbb{E}\|X\|_\infty \sim \sqrt{\log(p)}.$$

Logistic LASSO : Bernstein & complexity

Assume that $X \sim \mathcal{N}(0, I_p)$.

Bernstein condition satisfied with $\kappa = 1$.

$$\mathrm{comp}(B) = \mathbb{E} \sup_{\|t\|_1 \leq 1} \langle t, X \rangle = \mathbb{E}\|X\|_\infty \sim \sqrt{\log(p)}.$$

$$r(\rho) = \left[\frac{\rho A \mathrm{comp}(B)}{\sqrt{N}} \right]^{\frac{1}{2\kappa}} \sim \left(\frac{\rho \sqrt{\log(p)}}{\sqrt{N}} \right)^{\frac{1}{2}}.$$

Logistic LASSO : sparsity

Sparsity parameter

$$\Delta(\rho) = \inf_{h \in \rho S \cap r(2\rho) B_{L_2}} \sup_{f \in \partial \|\cdot\|_1 (f^*)} \langle h, f \rangle$$

Logistic LASSO : sparsity

Sparsity parameter

$$\Delta(\rho) = \inf_{h \in \rho S \cap r(2\rho) B_{L_2}} \sup_{f \in \partial \|\cdot\|_1 (f^*)} \langle h, f \rangle$$

$$f \in \partial \| \cdot \|_1 (f^*) \Leftrightarrow \begin{cases} f_j = +1 \text{ when } f_j^* > 0, \\ f_j = -1 \text{ when } f_j^* < 0, \\ f_j \in [-1, +1] \text{ when } f_j^* = 0. \end{cases}$$

Logistic LASSO : sparsity

Sparsity parameter

$$\Delta(\rho) = \inf_{h \in \rho S \cap r(2\rho) B_{L_2}} \sup_{f \in \partial \|\cdot\|_1 (f^*)} \langle h, f \rangle$$

$$f \in \partial \| \cdot \|_1(f^*) \Leftrightarrow \begin{cases} f_j = +1 \text{ when } f_j^* > 0, \\ f_j = -1 \text{ when } f_j^* < 0, \\ f_j \in [-1, +1] \text{ when } f_j^* = 0. \end{cases}$$

Choose h and define P as the projector on the sparsity pattern of f^*. Let s denote the sparsity of f^*.

$$\langle h, f \rangle = \langle (I-P)h, f \rangle + \langle Ph, f \rangle \underbrace{\geq \|(I-P)h\|_1 - \|Ph\|_1}_{f \text{ well chosen}} = \|h\|_1 - 2\|Ph\|_1 = \rho - 2\|Ph\|_1$$

Logistic LASSO : sparsity

Sparsity parameter

$$\Delta(\rho) = \inf_{h \in \rho S \cap r(2\rho) B_{L_2}} \ \sup_{f \in \partial \|\cdot\|_1 (f^*)} \langle h, f \rangle$$

$$f \in \partial \| \cdot \|_1 (f^*) \Leftrightarrow \begin{cases} f_j = +1 \text{ when } f_j^* > 0, \\ f_j = -1 \text{ when } f_j^* < 0, \\ f_j \in [-1, +1] \text{ when } f_j^* = 0. \end{cases}$$

Choose h and define P as the projector on the sparsity pattern of f^*. Let s denote the sparsity of f^*.

$$\langle h, f \rangle = \langle (I-P)h, f \rangle + \langle Ph, f \rangle \underbrace{\geq \|(I-P)h\|_1 - \|Ph\|_1}_{f \text{ well chosen}} = \|h\|_1 - 2\|Ph\|_1 = \rho - 2\|Ph\|_1$$

$$\|Ph\|_1 \leq \sqrt{s}\|Ph\|_2 \leq \sqrt{s}\|h\|_2 \leq \sqrt{s}r(2\rho)$$

Logistic LASSO : sparsity

Sparsity parameter

$$\Delta(\rho) = \inf_{h \in \rho S \cap r(2\rho) B_{L_2}} \sup_{f \in \partial \|\cdot\|_1 (f^*)} \langle h, f \rangle$$

$$f \in \partial \|\cdot\|_1 (f^*) \Leftrightarrow \begin{cases} f_j = +1 \text{ when } f_j^* > 0, \\ f_j = -1 \text{ when } f_j^* < 0, \\ f_j \in [-1, +1] \text{ when } f_j^* = 0. \end{cases}$$

Choose h and define P as the projector on the sparsity pattern of f^*. Let s denote the sparsity of f^*.

$$\langle h, f \rangle = \langle (I - P)h, f \rangle + \langle Ph, f \rangle \underbrace{\geq \|(I - P)h\|_1 - \|Ph\|_1}_{f \text{ well chosen}} = \|h\|_1 - 2\|Ph\|_1 = \rho - 2\|Ph\|_1$$

$$\|Ph\|_1 \leq \sqrt{s}\|Ph\|_2 \leq \sqrt{s}\|h\|_2 \leq \sqrt{s} r(2\rho)$$

Sparsity equation

$$\Delta(\rho) \geq (4/5)\rho \Leftrightarrow \rho \text{ such that } \frac{\rho}{r(2\rho)} \geq C\sqrt{s}.$$

Logistic LASSO : solving the sparsity equation

$$r(\rho) \sim \left(\frac{\rho\sqrt{\log(p)}}{\sqrt{N}} \right)^{\frac{1}{2}} .$$

$$C\sqrt{s} \leq \frac{\rho}{r(2\rho)} \sim \left(\frac{\rho\sqrt{N}}{\sqrt{\log(p)}} \right) .$$

Logistic LASSO : solving the sparsity equation

$$r(\rho) \sim \left(\frac{\rho\sqrt{\log(p)}}{\sqrt{N}} \right)^{\frac{1}{2}}.$$

$$C\sqrt{s} \leq \frac{\rho}{r(2\rho)} \sim \left(\frac{\rho\sqrt{N}}{\sqrt{\log(p)}} \right).$$

$$\rho^* \sim s\sqrt{\frac{\log(p)}{N}}.$$

Logistic LASSO : solving the sparsity equation

$$r(\rho) \sim \left(\frac{\rho\sqrt{\log(p)}}{\sqrt{N}} \right)^{\frac{1}{2}} .$$

$$C\sqrt{s} \leq \frac{\rho}{r(2\rho)} \sim \left(\frac{\rho\sqrt{N}}{\sqrt{\log(p)}} \right) .$$

$$\rho^* \sim s\sqrt{\frac{\log(p)}{N}} .$$

$$r(\rho^*) \sim \sqrt{\frac{s\log(p)}{N}} .$$

Logistic LASSO : conclusion

Theorem

Take $\lambda \sim \sqrt{\log(p)/N}$. Then with probability at least

$$1 - C \exp\left[-Cs \log(p)\right]$$

we have simultaneously

$$\|\hat{f} - f^*\|_1 \leq Cs\sqrt{\frac{\log(p)}{N}},$$

$$\|\hat{f} - f^*\|_2 \leq C\sqrt{\frac{s\log(p)}{N}},$$

$$R(\hat{f}) - R(f^*) \leq C\frac{s\log(p)}{N}.$$

The SLOPE penalty

	LASSO	SLOPE
$\|t\|$	$\sum_{i=1}^{p} \lvert t_i \rvert$	$\sum_{i=1}^{p} \sqrt{\log\left(\frac{ep}{i}\right)} \lvert t_{(i)} \rvert$
$\mathrm{comp}(B)$	$\sqrt{\log p}$	1
ρ^*	$\frac{s}{\sqrt{N}} \sqrt{\log p}$	$\frac{s}{\sqrt{N}} \log \frac{ep}{s}$
$r(\rho^*)$	$\frac{s}{N} \log p$	$\frac{s}{N} \log \frac{ep}{s}$

where $|t_{(1)}| \geq \cdots \geq |t_{(p)}|$.

Logistic SLOPE : conclusion

Theorem

Take $\lambda \sim 1/\sqrt{N}$. Then with probability at least

$$1 - C \exp\left[-Cs\log(\mathrm{e}p/s)\right]$$

we have simultaneously

$$\|\hat{f} - f^*\|_1 \leq Cs\sqrt{\frac{\log(\mathrm{e}p/s)}{N}},$$

$$\|\hat{f} - f^*\|_2 \leq C\sqrt{\frac{s\log(\mathrm{e}p/s)}{N}},$$

$$R(\hat{f}) - R(f^*) \leq C\frac{s\log(\mathrm{e}p/s)}{N}.$$

Matrix completion : context

$$E = F = \{\langle M, \cdot \rangle_F , M \in [-1, +1]^{m \times p}\} \text{ with } \| \cdot \| = \| \cdot \|_* .$$

Matrix completion via hinge loss + nuclear norm

$$\hat{f} \in \arg\min_{f \in F} \left[\frac{1}{N} \sum_{i=1}^{N} (1 - Y_i f(X_i))_+ + \lambda \|f\|_* \right] .$$

Matrix completion : context

$$E = F = \{\langle M, \cdot \rangle_F, M \in [-1,+1]^{m\times p}\} \text{ with } \|\cdot\| = \|\cdot\|_*.$$

Matrix completion via hinge loss + nuclear norm

$$\hat{f} \in \arg\min_{f\in F} \left[\frac{1}{N} \sum_{i=1}^{N} (1 - Y_i f(X_i))_+ + \lambda \|f\|_* \right].$$

Assume that X is uniformly distributed on $\{E_{j,k}\}$.

Matrix completion : Bernstein and complexity

Obvious that $f^*(E_{j,k}) = \mathrm{sign}(\langle E_{j,k}, M^* \rangle) = \mathrm{sign}(\eta(E_{j,k}))$. As soon as $|\eta(E_{j,k})| \geq \beta > 0$ then Bernstein satisfied with $\kappa = 1$.

Matrix completion : Bernstein and complexity

Obvious that $f^*(E_{j,k}) = \mathrm{sign}(\langle E_{j,k}, M^* \rangle) = \mathrm{sign}(\eta(E_{j,k}))$. As soon as $|\eta(E_{j,k})| \geq \beta > 0$ then Bernstein satisfied with $\kappa = 1$.

$$\mathrm{comp}(B) = \mathbb{E} \sup_{\|M\|_* \leq 1} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_i \langle M, X_i \rangle \right| = \mathbb{E} \sup_{\|M\|_* \leq 1} \left| \left\langle M, \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_i X_i \right\rangle \right|$$

$$= \mathbb{E} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_i X_i \right\|_{\mathrm{op}} \sim \sqrt{\frac{\log(m+p)}{\min(m,p)}}$$

thanks to "matrix Bernstein" inequality.

Matrix completion : Bernstein and complexity

Obvious that $f^*(E_{j,k}) = \mathrm{sign}(\langle E_{j,k}, M^* \rangle) = \mathrm{sign}(\eta(E_{j,k}))$. As soon as $|\eta(E_{j,k})| \geq \beta > 0$ then Bernstein satisfied with $\kappa = 1$.

$$\mathrm{comp}(B) = \mathbb{E} \sup_{\|M\|_* \leq 1} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_i \langle M, X_i \rangle \right| = \mathbb{E} \sup_{\|M\|_* \leq 1} \left| \left\langle M, \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_i X_i \right\rangle \right|$$

$$= \mathbb{E} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_i X_i \right\|_{\mathrm{op}} \sim \sqrt{\frac{\log(m+p)}{\min(m,p)}}$$

thanks to "matrix Bernstein" inequality.

$$r(\rho) \sim \left(\rho \sqrt{\frac{\log(m+p)}{N \min(m,p)}} \right)^{1/2}.$$

Matrix completion : sparsity

Sparsity equation

$$\Delta(\rho) \geq (4/5)\rho \Leftrightarrow \rho \text{ such that } \frac{\rho}{r(2\rho)} \geq C\sqrt{\mathrm{rank}(M^*)mp}.$$

Put $r = \mathrm{rank}(M^*)$.

Matrix completion : conclusion

Theorem

Take $\lambda \sim \sqrt{\log(m+p)/[N\min(m,p)]}$. Then with probability at least

$$1 - C\exp\left[-Cr(m+p)\log(m+p)\right]$$

we have simultaneously

$$\|\hat{f} - f^*\|_* \leq Cr\sqrt{\frac{\log(m+p)}{N\min(m,p)}},$$

$$\|\hat{f} - f^*\|_F \leq C\sqrt{\frac{r\max(m,p)\log(m+p)}{N}},$$

$$R(\hat{f}) - R(f^*) \leq C\frac{r\max(m,p)\log(m+p)}{N}.$$

P. Alquier, V. Cottet & G. Lecué (2017). Estimation Bounds and Sharp Oracle Inequalities of Regularized Procedures with Lipschitz Loss Functions. *Preprint arxiv :1702.01402*.

Jupyter notebooks :

`https://sites.google.com/site/vincentcottet/code`

Thank you !