Sharp oracle inequalities for non- convex loss

Sara van de Geer

December 22, 2017

Joint work with Andreas Elsener and Jana Jancová

Meeting in Mathematical Statistics

$\leq M$, $n \geq 1$, $M \geq 2$. *Let Assumption* $RE(s, 3)$ *…sso estimator* $\widehat{\beta}_L$ *defined by (7.2) with*

$$r = A\sigma\sqrt{\frac{\log M}{n}}$$

…robability at least $1 - M^{1-A^2/8}$, *we have*

$$|_1 \leq \frac{16A}{\kappa^2(s,3)} \sigma s \sqrt{\frac{\log M}{n}}$$

$$\beta^*)|_2^2 \leq \frac{16A^2}{\kappa^2(s,3)} \sigma^2 \ldots \log M,$$

$$\frac{64\phi_{\max}}{\kappa^2(s,3)} s.$$

…is satisfied, then with the same probability as … $1 \leq p \leq 2$ we have

$$\ldots + 3\sqrt{\frac{s}{m}}\Big\}^{2(p-1)} s \left(\frac{A\sigma}{\kappa^2(s,m,3)}\sqrt{\frac{\log M}{n}}\right)^p.$$

…milar to (7.7) and (7.8) can be deduced from

birthday Sasha

$\leq M$, $n \geq 1$, $M \geq 2$. *Let Assumption RE(s, 3)*
sso estimator $\widehat{\beta}_L$ *defined by* (7.2) *with*

$$r = A\sigma\sqrt{\frac{\log M}{n}}$$

robability at least $1 - M^{1-A^2/8}$, *we have*

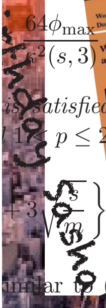
$$|_1 \leq \frac{16A}{\kappa^2(s,3)}\sigma s\sqrt{\frac{\log M}{n}}$$

$$\beta^*)|_2^2 \leq \frac{16A^2}{\kappa^2(s,3)}\sigma^2 \frac{s}{n}\log M,$$

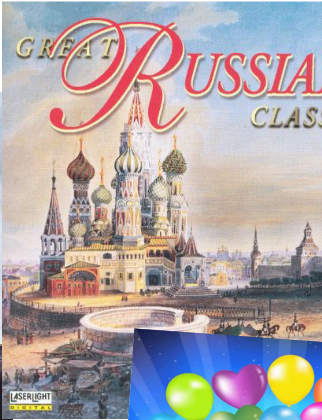
$$\frac{64\phi_{\max}}{\kappa^2(s,3)} s$$

is satisfied, then with the same probability as $1 \leq p \leq 2$ *we have*

$$+3\sqrt{\frac{s}{m}}\Big\}^{2(p-1)} s\left(\frac{A\sigma}{\kappa^2(s,m,3)}\right)$$



imilar to (7.7) and (7.8) can be deduced from



2) for any $p \in [1,\infty]$, $\varepsilon \le \varepsilon(\tau,q)$ *and any countable* $\mathrm{H} \subset \mathfrak{S}_d$

$$\mathbb{E}\Big\{\sup_{\vec h\in \mathrm{H}}\Big[\|\xi_{\vec h}\|_p - \widetilde{\Psi}_{\varepsilon,p}(\vec h)\Big]_+\Big\}^q \le \{C_3\varepsilon\}^q [illegible]$$

We will need also the following technical result.

Lemma 1. *For any* $d \ge 1$, $\varkappa \in (0,1/d)$, $\mathfrak{L} > 0$ *and* $\mathcal{A} \ge e^d$

(i) $\mathbb{H}_d(\varkappa,\mathfrak{L},\mathcal{A}) \subseteq \mathbb{H}_d\big(d\varkappa,\mathfrak{L}^d,\mathcal{A}\big)$;

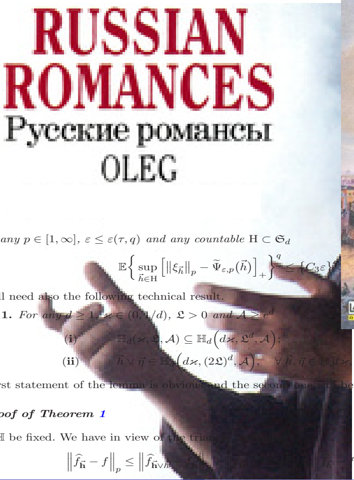
(ii) $\vec h \vee \vec\eta \in \mathbb{H}_d\big(d\varkappa,(2\mathfrak{L})^d,\mathcal{A}\big)$, $\quad \forall\, \vec h,\vec\eta \in \mathbb{H}_d(\varkappa,\mathfrak{L},\mathcal{A})$.

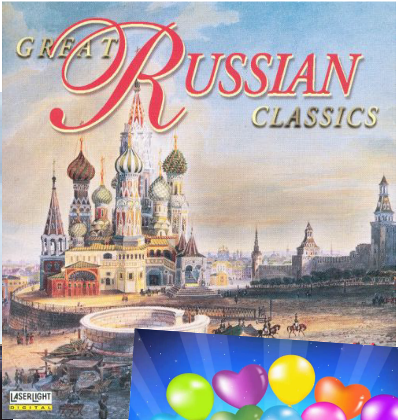
The first statement of the lemma is obvious and the second one will be proved in Appendix.

5.3. *Proof of Theorem 1*

Let $\vec h \in \mathbb{H}$ be fixed. We have in view of the tria

$$\Big\|\widehat f_{\vec{\mathbf h}} - f\Big\|_p \le \Big\|\widehat f_{\vec{\mathbf h}\vee\vec h} - [illegible] \Big\| [illegible] f_{\vec h} - f\Big\|_p. \qquad (5.1)$$





2) for any $p \in [1,\infty]$, $\varepsilon \le \varepsilon(\tau, q)$ and any countable $\mathrm{H} \subset \mathfrak{S}_d$

$$\mathbb{E}\Big\{ \sup_{\vec{h}\in \mathrm{H}} \Big[\|\xi_{\vec{h}}\|_p - \widetilde{\Psi}_{\varepsilon,p}(\vec{h}) \Big]_+ \Big\}^q \le \{C_3 \varepsilon\}$$

We will need also the following technical result.

Lemma 1. *For any $d \ge 1$, $\varkappa \in (0, 1/d)$, $\mathfrak{L} > 0$ and $\mathcal{A} \ge e^d$*

(i) $\mathbb{H}_d(\varkappa, \mathfrak{L}, \mathcal{A}) \subseteq \mathbb{H}_d\Big(d\varkappa, \mathfrak{L}^d, \mathcal{A}\Big)$;

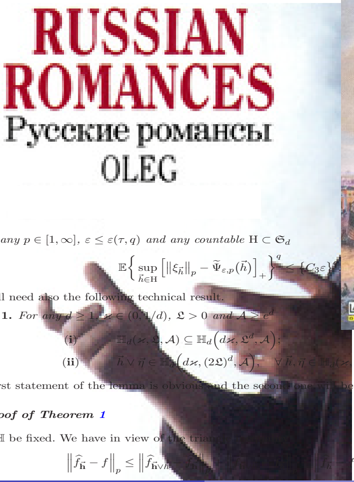
(ii) $\vec{h} \vee \vec{\eta} \in \mathbb{H}_d\Big(d\varkappa, (2\mathfrak{L})^d, \mathcal{A}\Big)$, $\quad \forall \vec{h}, \vec{\eta} \in \mathbb{H}_d(\varkappa, \mathfrak{L}, \mathcal{A})$.

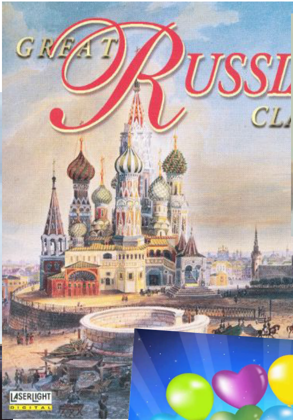
The first statement of the lemma is obvious and the second one will be proved in Appendix.

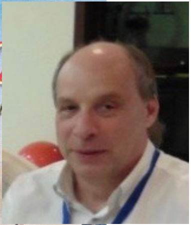
5.3. Proof of Theorem 1

Let $\vec{h} \in \mathbb{H}$ be fixed. We have in view of the tria

$$\Big\| \widehat{f}_{\vec{\mathbf{h}}} - f \Big\|_p \le \Big\| \widehat{f}_{\vec{\mathbf{h}} \vee \vec{h}} \quad f_{\vec{h}} - f \Big\|_p . \qquad (5.1)$$







2) for any $p \in [1,\infty]$, $\varepsilon \leq \varepsilon(\tau, q)$ *and any countable* $\mathrm{H} \subset \mathfrak{S}_d$

$$\mathbb{E}\Big\{ \sup_{\vec{h}\in\mathrm{H}} \Big[\|\xi_{\vec{h}}\|_p - \widetilde{\Psi}_{\varepsilon,p}(\vec{h}) \Big]_+ \Big\}^q \leq \{C_3\varepsilon\}^q$$

We will need also the following technical result.

Lemma 1. *For any* $d \geq 1$, $\varkappa \in (0, 1/d)$, $\mathfrak{L} > 0$ *and* $\mathcal{A} \geq e^d$

(i) $\mathbb{H}_d(\varkappa, \mathfrak{L}, \mathcal{A}) \subseteq \mathbb{H}_d\big(d\varkappa, \mathfrak{L}^d, \mathcal{A}\big)$;

(ii) $\vec{h} \vee \vec{\eta} \in \mathbb{H}_d\big(d\varkappa, (2\mathfrak{L})^d, \mathcal{A}\big), \quad \forall\, \vec{h}, \vec{\eta} \in \mathbb{H}_d(\varkappa, \mathfrak{L}, \mathcal{A})$.

The first statement of the lemma is obvious and the second one will be proved in Appendix.

5.3. Proof of Theorem 1

Let $\vec{h} \in \mathbb{H}$ be fixed. We have in view of the triangle

$$\big\|\widehat{f}_{\vec{\mathbf{h}}} - f\big\|_p \leq \big\|\widehat{f}_{\vec{\mathbf{h}}\vee\vec{h}} - \dots \big\|_p \dots \big\|f_{\vec{h}} - f\big\|_p. \qquad (5.1)$$

My favourite quote:

“You know adaptive estimators converge very fast if the function is very smooth (or has a prescribed complexity) but you can tell nothing about the estimated function itself”

Marc Hoffmann and Oleg Lepski (2002)

Aim in the rest of the talk

Show sharp oracle inequalities for

- ◦ global minimizers of convex but possibly non-differentiable loss
- ◦ stationary points of differentiable but possibly non-convex loss

Data:

- $X_1, \ldots, X_n$

Parameter space:

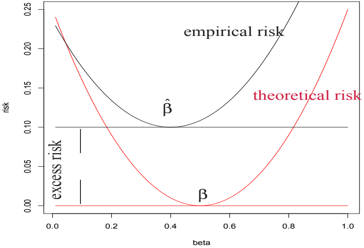
- $\mathcal{B} \subset \mathbb{R}^p$ convex.

Empirical (random) risk function:

- $\hat{R}_n(b)$, $b \in \mathcal{B}$.

Theoretical (nonrandom) risk function:

- $R(b)$, $b \in \mathcal{B}$.



<u>Aim</u>

Estimate

$$\beta^0 := \arg\min_{b \in \mathcal{B}} R(b).$$

We consider

$$\boxed{\beta^0 \text{ is high-dimensional: } p \gg n}$$

and

- $b \mapsto \hat{R}_n(b)$ is possibly not differentiable
- $b \mapsto \hat{R}_n(b)$ is possibly not convex and has multiple local minima
- $b \mapsto R(b)$ is convex

Example Least absolute deviations regression

Observed:

- $Y \in \mathbb{R}^n$
- $X \in \mathbb{R}^{n \times p}$

Empirical risk function

∘ $\hat{R}_n(b) := \frac{1}{n} \sum_{i=1}^{n} |Y_i - (Xb)_i|$

not really differentiable, sign-function not "smooth"

Example Linear regression with errors in variables

$$Y = X\beta^0 + \epsilon,$$

Observed:

- Y
- $Z = X + U$ with $U \perp X$, $\mathrm{cov}(U) := \Sigma_u$ known.

Let $\hat{\Sigma}_z := Z^T Z/n$.

We use

$$R_n(b) := Y^T Zb/n + b^T(\hat{\Sigma}_z - \Sigma_u)b.$$

$\hat{\Sigma}_z - \Sigma_u$ *is not necessarily positive semi-definite*
$\rightsquigarrow$ *possibly non-convex empirical risk*

Example Principal components

Observed:
• $X \in \mathbb{R}^{n\times p}$

Let $\hat{\Sigma} := X^T X / n$ and $\Sigma_0 := \mathbb{E}\hat{\Sigma}$

We aim at estimating the first eigenvector of Σ_0.

The risk function is for example

$$\hat{R}_n(b) := \|\hat{\Sigma} - bb^T\|_2^2$$

not convex

Example Estimation of an inverse Fisher information

Suppose $\ddot{R}(\beta^0)$ exists and we want to estimate its inverse

$$\ddot{R}^{-1}(\beta^0).$$

To estimate the first column of $\ddot{R}(\beta^0)$ use a node-wise Lasso

$$\min_{\gamma \in \mathbb{R}^p:\ \gamma_1 = 1} \gamma^T \hat{\ddot{R}}_n(\hat{\beta})\gamma + 2\lambda \|\gamma_{-1}\|_1 .$$

Since $\hat{\ddot{R}}_n(\hat{\beta})$ is not necessarily positive definite this is again a non-convex problem.

Our aim:
Extend the theory to sharp oracle inequalities when
- the empirical risk is not differentiable

or
- the empirical risk is not convex

Related work:

Po-Ling Loh and Martin Wainwright (2014, 2015)

Song Mei, Yu Bai, and Andrea Montanari (2016)

New: sharp oracle inequalities

Main idea from:
Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion,
Koltchinskii, V. and Lounici, K. and Tsybakov, A.B.

The estimator

To deal with $\boxed{\beta \text{ high-dimensional}}$ we consider a norm Ω on $\mathbb{R}^p$.
The regularized empirical risk is

$$\hat{R}_n(b) + \lambda\Omega(b)$$

where $\lambda > 0$ is a tuning parameter.

Definition (argmin) *Let*

$$\hat{\beta} := \hat{\beta}_{\mathrm{argmin}} := \min_{b \in \mathcal{B}} \left\{ \hat{R}_n(b) + \lambda \Omega(b) \right\}$$

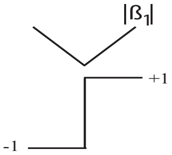
Sub-differential:

The sub-differential of Ω is

$$\partial \Omega(b) := \left\{ z \in \mathbb{R}^p : \ \Omega^*(z) \le 1, \ z^T b = \Omega(b) \right\}$$

Example: $\Omega = \| \cdot \|_1$

$\partial \|b\|_1 =$
$\{z \in \mathbb{R}^p : \|z\|_\infty \le 1,$
$z_j = \mathrm{sign}(\beta_j), \ \beta_j \neq 0\}$



subdifferential calculus

Suppose $\partial \hat{R}_n(b)/\partial b := \dot{\hat{R}}_n^T(b)$ exists.

Definition (stationary) *Let $\hat{\beta} := \hat{\beta}_{\rm stationary}$ be a solution of the KKT conditions*

$$\dot{\hat{R}}_n^T(\hat{\beta}) + \lambda \hat{z} = 0,\ \hat{z} \in \partial\Omega(\hat{\beta}).$$

Definition (semi stationary) *Let $\hat{\beta} := \hat{\beta}_{\rm semi\ stationary}$ satisfy*

$$\dot{\hat{R}}_n^T(\hat{\beta})(\hat{\beta} - \beta) + \lambda\Omega(\hat{\beta}) - \lambda\Omega(\beta) \le 0.$$

Here, and throughout, $\beta \in \mathcal{B}$ is fixed (not necessarily $\beta = \beta^0$).

Note

$$\hat{\beta} = \hat{\beta}_{\text{argmin}} \;\Rightarrow\; \hat{\beta} = \hat{\beta}_{\text{semi stationary}}$$

$$\hat{\beta} = \hat{\beta}_{\text{stationary}} \Rightarrow \; \hat{\beta} = \hat{\beta}_{\text{semi stationary}}$$

Let τ be some semi-norm on $\mathbb{R}^p$ and $G : [0, \infty) \to [0, \infty)$ be an increasing strictly convex function with $G(0) = 0$.

Definition *We say that strict convexity holds if* $\forall\ 0 \le t \le 1$ *sufficiently small*

$$R\Big((1-t)b + t\beta\Big) \le (1-t)R(b) + tR(\beta) - t(1-t)G\Big(\tau(\beta - b)\Big)$$

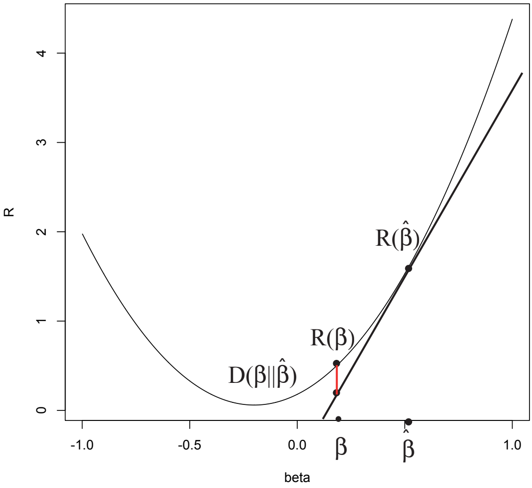
for all $b \in \mathcal{B}$*.*

Definition *We say that the Bregman condition holds if*

$$R(\beta) - R(b) \ge \dot{R}^T(b)(\beta - b) + G\Big(\tau(\beta - b)\Big)$$

for all $b \in \mathcal{B}$*.*

Note G-convexity $\Rightarrow$ G-margin condition
Note G is a convex lower bound for the "Bregman divergence"



Bregman divergence

Results when $\Omega = \|\cdot\|_1$

We first consider the ℓ_1-penalty.

Let for $S \subset \{1, \ldots, p\}$ and $b \in \mathbb{R}^p$,

$$b_S = \{b_j \mathrm{l}\{j \in S\}\}, \ b_{-S} = \{b_j \mathrm{l}\{j \notin S\}\}$$

and

$$S_b := \{j : \ b_j \neq 0\}.$$

Example: $p = 7$, $|S| = 3$, $S = \{2, 3, 7\}$

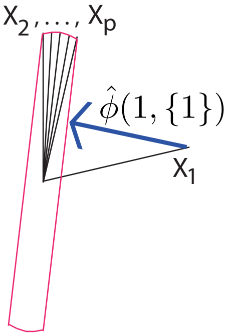
$$b = \begin{pmatrix} * \\ * \\ * \\ * \\ * \\ * \\ * \end{pmatrix} \qquad b_S = \begin{pmatrix} 0 \\ * \\ * \\ 0 \\ 0 \\ 0 \\ * \end{pmatrix} \qquad b_{-S} = \begin{pmatrix} * \\ 0 \\ 0 \\ * \\ * \\ * \\ 0 \end{pmatrix}$$

Definition *The <u>effective sparsity</u> at S with stretching constant L is*

$$\Gamma^2(L, S) = \max\left\{ \frac{\|b\|_1^2}{\tau^2(b)} : \|b_{-S}\|_1 \le L\|b_S\|_1 \right\}.$$

Remark: Think of $\Gamma^2(L, S)$ as being of the flavour $\asymp |S|$

We have
$\hat{\Gamma}^2(L, S) = |S|/\hat{\phi}^2(L, S)$
where $\hat{\phi}^2(L, S)$
="compatibility constant"
$\approx$ "restricted eigenvalue"



Theorem argmin (No differentiablity assumed)

Let $\hat{\beta} = \hat{\beta}_{\text{argmin}}$.

Assume strict convexity with $G(u) = u^2/2$. *Let for appropriate fixed* $0 < t < 1$

$$\lambda_0 := \lambda_{\text{argmin}} \geq \frac{\left| [\hat{R}_n - R]\Big((1-t)\hat{\beta} + t\beta \Big) - [\hat{R}_n - R](\hat{\beta}) \right|}{t \|\hat{\beta} - \beta\|_1 + 1/n}.$$

We refer to λ_0 *as the noise level.*

For $\lambda > \lambda_0$ *we have*

$$R(\hat{\beta}) \leq R(\beta) + (\lambda + \lambda_0)^2 \Gamma^2(L, S_\beta)/2$$

with $L := (\lambda + \lambda_0)/(\lambda - \lambda_0)$.

Flavour of Theorem argmin (No differentiablity assumed)

Let $\hat{\beta} = \hat{\beta}_{\text{argmin}}$.

Assume strict convexity with $G(u) = u^2/2$. *Let for appropriate fixed* $0 < t < 1$

$$\lambda_0 := \lambda_{\text{argmin}} \geq \frac{\left| [\hat{R}_n - R]\Big((1-t)\hat{\beta} + t\beta\Big) - [\hat{R}_n - R](\hat{\beta}) \right|}{t\|\hat{\beta} - \beta\|_1 + 1/n}.$$

We refer to λ_0 *as the* *noise level.*

Then $\lambda_0 \asymp \sqrt{\log p/n}$ *and for* $\lambda \asymp \lambda_0$ *we have*

$$R(\hat{\beta}) - R(\beta) \asymp \lambda^2 |S_\beta|$$

with high probability.

Theorem semi stationary (Differentiablity assumed)

Consider $\hat{\beta} = \hat{\beta}_{\text{semi stationary}}$.

Assume the Bregman condition with $G(u) = u^2/2$.

Let

$$\lambda_0 := \lambda_{\text{semi stationary}} \geq \frac{\left|(\dot{\hat{R}}_n - \dot{R})^T(\hat{\beta} - \beta)\right|}{\|\hat{\beta} - \beta\|_1 + 1/n}.$$

We refer to λ_0 *as the* *noise level.*

For $\lambda > \lambda_0$ *we have*

$$R(\hat{\beta}) \leq R(\beta) + (\lambda + \lambda_0)^2 \Gamma^2(L, S_\beta)/2$$

with $L := (\lambda + \lambda_0)/(\lambda - \lambda_0)$.

Remark Both theorems have the same flavour.

Remark For general G we get in both theorems

$$R(\hat{\beta}) \leq R(\beta) + H\Big((\lambda + \lambda_0)\Gamma(L, S_\beta)\Big).$$

where H is the convex conjugate of G.

Example $G(u) = u^2/2 \Rightarrow H(v) = v^2/2$.

Remark

Determining the noise level λ_0 is the random part of the problem.

The noise level λ_0 should be such that the inequality in the theorems holds with large probability.

We call this the empirical process condition.

Example: generalized linear models ...

Generalized linear model

Suppose

$$\hat{R}_n(b) = \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i, (Xb)_i)$$

with a normalized X.

Assumption (ρ Lipschitz) *$z \mapsto \rho(\cdot, z)$ is Lipschitz.*

Let $\dot{\rho}(\cdot, z) := d\rho(\cdot, z)/dz$.

Assumption ($\dot{\rho}$ Lipschitz) *$z \mapsto \dot{\rho}(\cdot, z)$ is Lipschitz.*

concentration, contraction, peeling .... $\rightsquigarrow$
Assumption (ρ-Lipschitz) $\Rightarrow \lambda_{\rm argmin} \asymp \sqrt{\frac{\log p}{n}}$
Assumption ($\dot{\rho}$-Lipschitz) $\Rightarrow \lambda_{\rm semi\ stationary} \asymp \sqrt{\frac{\log p}{n}}$

Thus, the empirical process condition $\rightsquigarrow$ the $\asymp$ usual value for λ_0

Conclusion

∘ Assumption (ρ-Lipschitz)

$\Rightarrow$ sharp oracle inequalities for the global minimizer

$\rightsquigarrow$ useful for the convex non-differentiable case

∘ Assumption ($\dot{\rho}$-Lipschitz)

$\Rightarrow$ sharp oracle inequalities for (semi) stationary points

$\rightsquigarrow$ useful for non convex case

Note

In both cases we assume $\beta \mapsto R(\beta)$ to be convex
The non convexity is about $\beta \mapsto \hat{R}_n(\beta)$

An example of a generalized linear model

Non-differentiable case

Example <u>Least absolute deviations</u>

Observed

- $Y \in \mathbb{R}^n$
- $X \in \mathbb{R}^{n \times p}$

Model

$$Y = X\beta^0 + \epsilon$$

with

$\epsilon_1, \ldots, \epsilon_n$ i.i.d. with density f_0

We take

$$\rho(y, (xb)) = |y - xb|.$$

Assumption (ρ Lipschitz) holds.

But

$$\dot{\rho}(y, z) = -\text{sign}(y - z).$$

Assumption ($\dot{\rho}$ Lipschitz) does not hold.

We have:

$$\begin{matrix} \mathcal{B} \subset \{b : \ \|Xb\|_\infty \le \text{const.}\} \\ \text{fixed design (say)} + \text{cond.}^{\text{s}} \ f_0 \end{matrix} \quad \Rightarrow \text{(strictconvexity)}$$

with $G(u) = \text{const. } u^2$ and $\tau^2(b) = \|Xb\|_2^2/n$.

It follows that with high probability

$$R(\hat{\beta}) \le R(\beta) + \mathcal{O}\left(\frac{\log p}{n}\right)\frac{|S_\beta|}{\phi^2(S_\beta, L)},$$

where

$$\phi^2(L, S) := |S|/\Gamma^2(L, S) = \min\left\{\|Xb\|_2^2/n : \ \|b_{-S}\|_1 \le L\|b_S\|_1\right\}$$

is the compatibility constant ($\approx$ restricted eigenvalue).

Beyond generalized linear models: some examples

Example Sparse principal components

Observed:

- $X \in \mathbb{R}^{n \times p}$

Let $\hat{\Sigma} := X^T X / n$ and $\Sigma_0 := \mathbb{E}\hat{\Sigma}$

We aim at estimating the first scaled eigenvector β^0 of Σ_0:

$$\beta^0 = \arg\min_b \|\Sigma_0 - bb^T\|_2^2.$$

Assumption (subGaus)
X has i.i.d. sub-Gaussian rows.

Assumption (gap)
There is a gap ~ 1 *between the first and second eigenvalue of* Σ_0*.*

Assumption (sparse)
$s_0 := |S_{\beta^0}| = o(\sqrt{n/\log p})$ *(or a "weak sparsity" version)*

First step: localizing

$$\hat{Z} := \arg \min_{\mathrm{trace}(Z)=1, 0 \le Z \le I} \left\{ -\mathrm{trace}(\hat{\Sigma} Z) + \lambda \| Z \|_1 \right\}.$$

[d'Aspremont, El Ghaoui, Jordan, Lanckriet (2007)]
[Vu, Cho, Lei, Rohe (2013)]

$\rightsquigarrow \hat{\beta}_{\mathrm{init}}$ with $\| \hat{\beta}_{\mathrm{init}} - \beta^0 \|_2 = o_{\mathbb{P}}(1)$.

Second step: nonconvex loss

We now let

$$\hat{R}_n(b) := \|\hat{\Sigma} - bb^T\|_2^2/4$$

and so

$$\dot{\hat{R}}_n(b) = -\hat{\Sigma}b + \|b\|^2 b.$$

We let $\hat{\beta} = \hat{\beta}_{\text{semi stationary}} \in \mathcal{B} := \{b : \ \|b - \hat{\beta}_{\text{init}}\|_2 \leq \eta\}$:

$$\dot{\hat{R}}_n^T(\hat{\beta})(\hat{\beta} - \beta) + \lambda\|\hat{\beta}\| - \lambda\|\beta\|_1 \leq 0.$$

Note

$$(\dot{\hat{R}}_n - \dot{R})^T(\hat{\beta})(\hat{\beta} - \beta) = -\hat{\beta}^T(\hat{\Sigma} - \Sigma_0)(\hat{\beta} - \beta)$$

Using assumption (subGaus) we get

$$\left|(\dot{\hat{R}}_n - \dot{R})^T(\hat{\beta})(\hat{\beta} - \beta)\right| = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n}}\right)\left(\|\hat{\beta} - \beta\|_1 + \frac{1}{n}\right).$$

We obtain $\lambda_0 \asymp \sqrt{\log p/n}$ in Theorem semi stationary.

The Bregman condition holds with $G(u) = \mathrm{const}.u^2$ and $\tau(b) = \|b\|_2$.

The effective sparsity is thus $\Gamma^2(L, S) \sim |S|$.

It follows that

$$R(\hat{\beta}) \leq R(\beta) + \mathcal{O}\left(\frac{\log p}{n}\right)|S_\beta|.$$

De-biasing in sparse PCA

Suppose the parameter of interest is β_1^0.
We have

$$\ddot{\hat{R}}_n(\hat{\beta}) = -\hat{\Sigma} - \|\hat{\beta}\|_2^2 I + 2\hat{\beta}\hat{\beta}^T.$$

We obtain the first column of the surrogate inverse of $\ddot{\hat{R}}_n(\hat{\beta})$ by doing a "node-wise" Lasso:

$$\begin{aligned}
\hat{\gamma}_{\text{argmin}} &:= \arg \min_{\gamma^T = (1, \gamma_2, \ldots, \gamma_p)} \left\{ \gamma^T \ddot{\hat{R}}_n(\hat{\beta}) \gamma + 2\lambda_1 \|\gamma\|_1 \right\} \\
\hat{\gamma} &:= \hat{\gamma}_{\text{stationary}} : \left(\ddot{\hat{R}}_n(\hat{\beta}) \right)_{-1,\cdot} \hat{\gamma} + \lambda \hat{z}_{-1} = 0 \\
\hat{\Theta}_{1,1}^{-1} &:= \hat{\gamma}^T \ddot{\hat{R}}_n(\hat{\beta}) \hat{\gamma}, \\
\hat{\Theta}_1 &:= \hat{\gamma} \hat{\Theta}_{1,1}.
\end{aligned}$$

Note: $\ddot{\hat{R}}_n(\hat{\beta})$ is not necessarily p.s.d. $\rightsquigarrow$ non-convex problem.

The de-biased estimator is

$$\hat{b}_1 := \hat{\beta}_1 - \hat{\Theta}_1^T \Big(\underbrace{\|\hat{\beta}\|_2^2 \hat{\beta} - \hat{\Sigma}\hat{\beta}}_{\dot{\hat{R}}_n(\hat{\beta})} \Big).$$

Lemma *Assume*

$\circ\ \lambda \asymp \sqrt{\log p/n}$
$\circ\ \lambda_1 \asymp \sqrt{\log p/n}$
$\circ\ s_0 = o(\sqrt{n}/\log p)$
$\circ\ s_1 = o(\sqrt{n}/\log p)$

Then

$$\sqrt{n}(\hat{b}_1 - \beta_1^0) \to \mathcal{N}(0, \sigma_1^2)$$

where

$$\sigma_1^2 = n\mathrm{var}(\Theta_1^{0\,T} \hat{\Sigma} \beta^0).$$

Results for general Ω

Let Ω be a norm on $\mathbb{R}^p$.
Recall

$$\hat{\beta}_{\text{argmin}} := \arg\min_{b \in \mathcal{B}} \left\{ \hat{R}_n(b) + \lambda \Omega(b) \right\},$$

etc.
For $\hat{\beta} = \hat{\beta}_{\text{argmin}}$ we assume strict convexity.
For $\hat{\beta} = \hat{\beta}_{\text{semi stationary}}$ we assume the Bregman condition.
Definition *The triangle property holds if* $\forall\ b$

$$\Omega(\beta) - \Omega(b) \leq \Omega^+(\beta - b) - \Omega^-(b).$$

We then write $\underline{\Omega} := \Omega^+ + \Omega^-$.
Definition *The effective sparsity is*

$$\Gamma^2(L) := \max\{\tau^2(b) :\ \Omega^-(b) \leq L,\ \Omega^+(b) = 1\}.$$

Theorem

Let

$$\lambda_{\text{argmin}} \geq \frac{\left|(\hat{R}_n - R)((1-t)\hat{\beta} + t\beta) - (\hat{R}_n - R)(\hat{\beta})\right|}{t\underline{\Omega}(\hat{\beta} - \beta) + 1/n}$$

$$\lambda_{\text{semi stationary}} \geq \frac{\left|(\dot{\hat{R}}_n - \dot{R})^T(\hat{\beta} - \beta)\right|}{\underline{\Omega}(\hat{\beta} - \beta) + 1/n}.$$

Define for appropriate $\lambda_0 \in \{\lambda_{\text{argmin}}, \lambda_{\text{semi stationary}}\}$

$$L = \frac{\lambda + \lambda_0}{\lambda - \lambda_0}.$$

Then for appropriate $\hat{\beta} \in \{\hat{\beta}_{\text{argmax}}, \hat{\beta}_{\text{semi stationary}}\}$

$$R(\hat{\beta}) \leq R(\beta) + H\Big((\lambda + \lambda_0)\Gamma(L)\Big)$$

where H is the convex conjugate of G.

Examples of norms used

ℓ_1-norm: $\Omega(b) = \|b\|_1 =: \sum_{j=1}^p |b_j|$

Oscar: given $\tilde{\lambda} > 0$

$$\Omega(b) := \sum_{j=1}^{p} (\tilde{\lambda}(j-1)+1)|b|_{(j)} \quad \text{where } |b|_{(1)} \geq \cdots \geq |b|_{(p)}$$

[Bondell and Reich 2008]

sorted ℓ_1-norm: given $\lambda_1 \geq \cdots \geq \lambda_p > 0$,

$$\Omega(b) := \sum_{j=1}^{p} \lambda_j |b|_{(j)} \qquad \text{where } |b|_{(1)} \geq \cdots \geq |b|_{(p)}$$

[Bogdan et al. 2013]

Examples of norms used

ℓ_1-norm: $\Omega(b) = \|b\|_1 =: \sum_{j=1}^{p} |b_j|$

Oscar: given $\tilde{\lambda} > 0$

$$\Omega(b) := \sum_{j=1}^{p} (\tilde{\lambda}(j-1)+1)|b|_{(j)} \quad \text{where } |b|_{(1)} \geq \cdots \geq |b|_{(p)}$$

[Bondell and Reich 2008]

sorted ℓ_1-norm: given $\lambda_1 \geq \cdots \geq \lambda_p > 0$,

$$\Omega(b) := \sum_{j=1}^{p} \lambda_j |b|_{(j)} \qquad \text{where } |b|_{(1)} \geq \cdots \geq |b|_{(p)}$$

[Bogdan et al. 2013]

Examples of norms used

ℓ_1-norm: $\Omega(b) = \|b\|_1 =: \sum_{j=1}^{p} |b_j|$

Oscar: given $\tilde{\lambda} > 0$

$$\Omega(b) := \sum_{j=1}^{p} (\tilde{\lambda}(j-1)+1)|b|_{(j)} \quad \text{where } |b|_{(1)} \geq \cdots \geq |b|_{(p)}$$

[Bondell and Reich 2008]

sorted ℓ_1-norm: given $\lambda_1 \geq \cdots \geq \lambda_p > 0$,

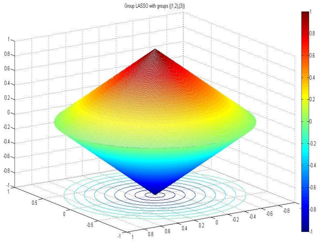
$$\Omega(b) := \sum_{j=1}^{p} \lambda_j |b|_{(j)} \qquad \text{where } |b|_{(1)} \geq \cdots \geq |b|_{(p)}$$

[Bogdan et al. 2013]

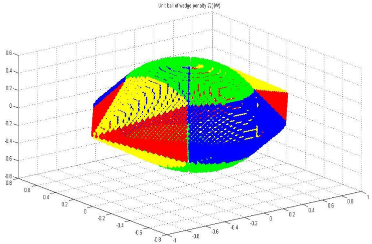
norms generated from cones:

$$\Omega(b) := \min_{a \in \mathcal{A}} \frac{1}{2} \sum_{j=1}^{p} \left[\frac{b_j^2}{a_j} + a_j \right], \mathcal{A} \subset \mathbb{R}^p_+ \text{ a convex cone.}$$

[Micchelli et al. 2010] [Jenatton et al. 2011] [Bach et al. 2012]



unit ball for group Lasso norm



unit ball for wedge norm
$\mathcal{A} = \{a : \ a_1 \geq a_2 \geq \cdots\}$

nuclear norm for matrices: $B \in \mathbb{R}^{p_1 \times p_2}$,

$$\Omega(B) := \|B\|_{\text{nuclear}} := \text{trace}(\sqrt{B^T B})$$

nuclear norm for tensors: $B \in \mathbb{R}^{p_1 \times p_2 \times p_3}$,

$\Omega(B) :=$ dual norm of Ω_*
where

$$\Omega_*(W) := \max_{\|u_1\|_2 = \|u_2\|_2 = \|u_3\|_2 = 1} \text{trace}(W^T u_1 \otimes u_2 \otimes u_3), \; W \in \mathbb{R}^{p_1 \times p_2 \times p_3}.$$

[Yuan and Zhang 2014]

nuclear norm for matrices: $B \in \mathbb{R}^{p_1 \times p_2}$,

$$\Omega(B) := \|B\|_{\text{nuclear}} := \text{trace}(\sqrt{B^T B})$$

nuclear norm for tensors: $B \in \mathbb{R}^{p_1 \times p_2 \times p_3}$,

$\Omega(B) :=$ dual norm of Ω_*
where

$$\Omega_*(W) := \max_{\|u_1\|_2 = \|u_2\|_2 = \|u_3\|_2 = 1} \text{trace}(W^T u_1 \otimes u_2 \otimes u_3), \ W \in \mathbb{R}^{p_1 \times p_2 \times p_3}.$$

[Yuan and Zhang 2014]

Example Matrix completion using robust loss
[Elsener and vdG, 2016]
Let X_i be a mask with a "1" at a random entry.

$$X_i := \begin{pmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{pmatrix}$$

$$\hat{R}_n(B) := \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i - \mathrm{trace}(X_i B))$$

where

- $\rho = \rho_{\mathrm{Huber}}$ or $\rho = \rho_{\mathrm{LAD}}$
- $B \in \mathcal{B} := \{B \in \mathbb{R}^{p_1 \times p_2} : \ \|B\|_\infty \leq \eta\}$ for some given η

Let $\Omega := \|\cdot\|_{\text{nuclear}}$.

Dual norm: use symmetrization, contraction, concentration ... more complicated due to non-linear random term, but doable

Margin semi-norm:
$\tau^2(B) = \|B\|_2^2/(p_1 p_2)$
Margin curvature:
$G(u) = u^2/(2cp_1 p_2)$
Effective sparsity:
$\Gamma^2(L) = 3s_B$, $s_B := \text{rank}(B)$.

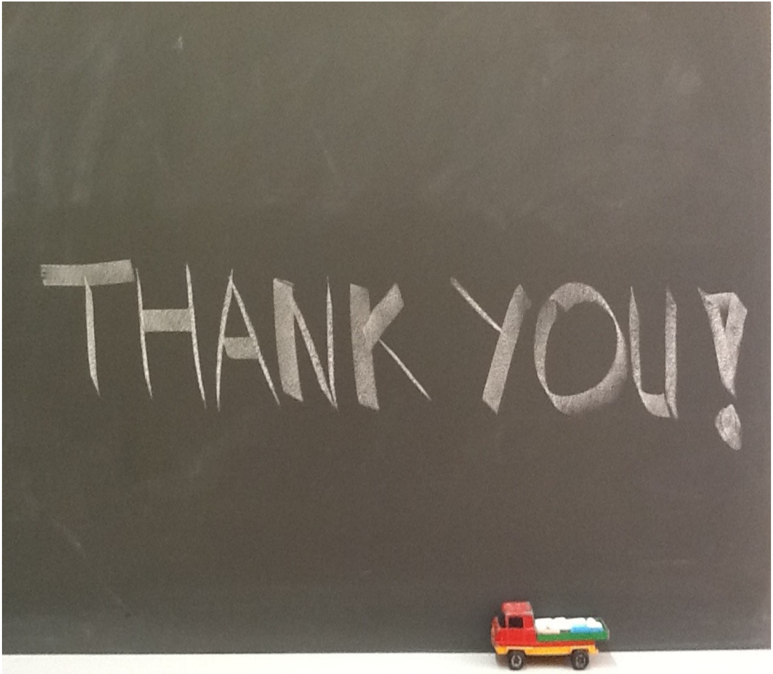
From Theorem argmin

for $p_1 \geq p_2$

and $\lambda = C_0 \frac{1}{\sqrt{np_2}}(\sqrt{\log p_1 + \log(1/\alpha)/p_1}$,

with probability at least $1 - \alpha$

$$R(\hat{B}) \leq R(B) + C \times \left(\frac{p_1 s_B \log(p_1)}{n} \right).$$


THANK YOU!