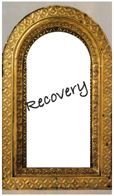
High-dimensional statistics: a triptych

Sara van de Geer

Panel I: Oracle inequalities

Panel II: Asymptotic normality and CRLB's



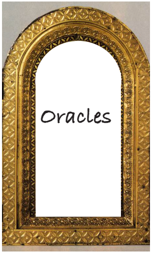
Panel III : Lower bounds for restricted eigenvalues

High-dimensional statistics: a triptych

Sara van de Geer

July 12, 2016

Panel I: Oracle inequalities



Joint work with:
Andreas Elsener, Alan Muro, Jana Janková, Benjamin Stucky

Norm-penalized empirical risk minimization

Examples of risk functions and norms

Main conditions: triangle property, margin curvature, effective sparsity

Oracle inequality

Illustrations:

Lasso

Matrix completion

Sparse PCA

Norm-penalized empirical risk minimization

Examples of risk functions and norms

Main conditions: triangle property, margin curvature, effective sparsity

Oracle inequality

Illustrations:

Lasso
Matrix completion
Sparse PCA

Goal

- derive sharp oracle inequalities
  for norm-penalized empirical risk minimizers
  - oracle: an estimator that knows the "true" sparsity
  - oracle inequalities: adaptation to unknown sparsity
  - sharp: learning point of view
- show the main ingredients making such results possible

sparsity:
roughly speaking: the number of active parameters is small

active parameter: roughly speaking: it is $\neq 0$

Regularized empirical risk minimization

$$\hat{\beta} := \arg \min_{\beta \in \mathcal{B} \subset \mathbb{R}^p} \Bigg\{ \underbrace{\hat{R}_n(\beta)}_{\text{empirical risk}} + \underbrace{\text{pen}(\beta)}_{\text{regularization penalty}} \Bigg\}$$

Regularized empirical risk minimization

$$\hat{\beta} := \arg \min_{\beta \in \mathcal{B} \subset \mathbb{R}^p} \left\{ \underbrace{\hat{R}_n(\beta)}_{\text{empirical risk}} + \underbrace{\text{pen}(\beta)}_{\text{regularization penalty}} \right\}$$

Example: Lasso

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

where

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

$n :=$ number of observations, $p :=$ number of parameters.

High-dimensional: $p \gg n$

target $\beta^0 \in \mathbb{R}^p$

data $X_1, \ldots, X_n$

number of parameters

$$p$$

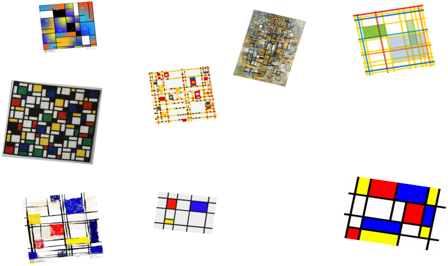
number of observations

$$n$$

high-dimensional statistics:

$$p \gg n$$

DATA



$$X_1, \ldots, X_n$$

$$\downarrow$$

$$\hat{\beta} \in \mathbb{R}^p$$

Classical least squares when $p < n$

Model:

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

$Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\mathrm{rank}(X) = p$ $(< n)$,

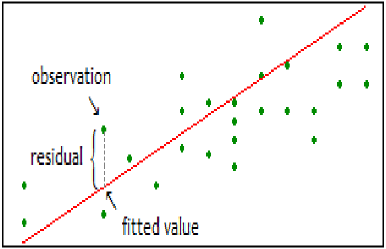
$\epsilon \sim \mathcal{N}_n(0, \sigma_0^2 I)$, $\sigma_0^2 := 1$.

Least squares estimator:

$$\hat{\beta}_{\mathrm{LS}} = (X^T X)^{-1} X^T Y.$$

Excess risk:

$$\mathcal{E}(\beta) := \|X(\beta - \beta^0)\|_2^2/n$$
$$= \mathbb{E}\|Y - X\beta\|_2^2/n - \underset{\substack{\uparrow \\ =1}}{\sigma_0^2}$$



We have

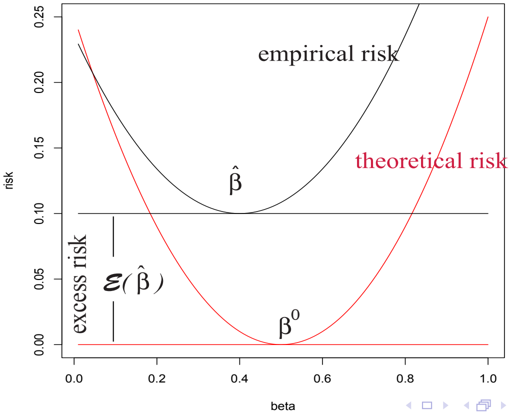
$$\mathbb{E}\mathcal{E}(\hat{\beta}_{\text{LS}}) = \frac{p}{n} = \frac{\text{number of parameters}}{\text{number of observations}}.$$

Aim in high-dimensional statistics

theoretical risk: $R(\beta) := \mathbb{E}\hat{R}_n(\beta)$

target: $\beta^0 := \arg\min_{\beta \in \mathcal{B}} R(\beta)$ ($\approx$ "true" parameter)

excess risk: $\mathcal{E}(\beta) := R(\beta) - R(\beta^0)$



Aim is oracle result:

with high probability and up to $\log p$ terms

$$\mathcal{E}(\hat{\beta}) \le \frac{s_0}{n} = \frac{\text{number of active parameters}}{\text{number of observations}}$$

where

$$s_0 = |S_0|$$

with

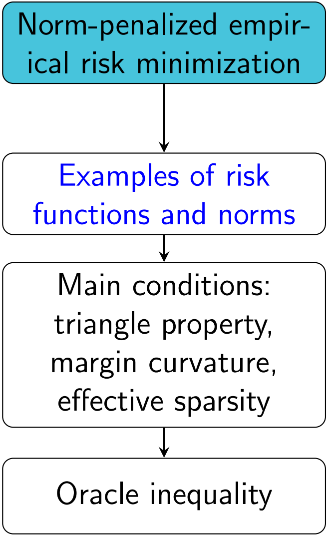
$$S_0 := \{j : \ \beta_j^0 \neq 0\} \text{ the active set of } \beta^0.$$

Norm-regularized empirical risk minimization

$$\hat{\beta} := \arg \min_{\beta \in \mathcal{B} \subset \mathbb{R}^p} \left\{ \underbrace{\hat{R}_n(\beta)}_{\text{empirical risk}} + \underbrace{\lambda \Omega(\beta)}_{\text{regularization penalty}} \right\}$$

where

- ∘ $\mathcal{B} \subset \mathbb{R}^p$ is convex
- ∘ Ω is a given norm on $\mathbb{R}^p$,
- ∘ $\lambda > 0$ is a tuning parameter



Norm-penalized empirical risk minimization

Examples of risk functions and norms

Main conditions: triangle property, margin curvature, effective sparsity

Oracle inequality

Illustations:

Lasso
Matrix completion
Sparse PCA

Examples of empirical risk functions

First some terminology

We have

$$\hat{R}_n(\beta) = \underbrace{\hat{R}_n(\beta) - \overset{\overset{:=\mathbf{E}\hat{R}_n(\beta)}{\downarrow}}{R(\beta)}}_{\text{“random part”}} + \underbrace{R(\beta)}_{\text{deterministic}}$$

Definition

We say that the “random part” is ***linear*** *if for a random vector* $\begin{pmatrix} W_0 \\ W \end{pmatrix} \in \mathbb{R}^{p+1}$

$$R_n(\beta) - R(\beta) = W_0 - \beta^T W \ \ \forall \ \beta$$

.

Example: Least squares with fixed design
Let $\hat{\Sigma} := X^T X/n$ be the Gram matrix and

$$\begin{aligned}\hat{R}_n(\beta) &:= \|Y - X\beta\|_2^2/(2n) \\ &= \frac{1}{2}\|Y\|_2^2/n - \beta^T X^T Y/n + \frac{1}{2}\beta^T \hat{\Sigma}\beta,\end{aligned}$$

The random part is

$$W = X^T \epsilon/n, \ \underset{\substack{\uparrow \\ \text{noise}}}{\epsilon} = Y - \mathbb{E}Y$$

Example: Least squares with fixed design
Let $\hat{\Sigma} := X^T X / n$ be the Gram matrix and

$$\begin{aligned} \hat{R}_n(\beta) &:= \| Y - X\beta \|_2^2 / (2n) \\ &= \frac{1}{2} \| Y \|_2^2 / n - \beta^T X^T Y / n + \frac{1}{2} \beta^T \hat{\Sigma} \beta, \end{aligned}$$

The random part is

$$W = X^T \epsilon / n, \ \underset{\substack{\uparrow \\ \text{noise}}}{\epsilon} = Y - \mathbb{E} Y$$

Example: Linearized least squares with known Σ_0
Let $\Sigma_0 := \mathbb{E} X^T X / n$ and

$$\hat{R}_n(\beta) := -\beta^T X^T Y / n + \beta^T \Sigma_0 \beta / 2$$

Then

$$W = (X^T Y - \mathbb{E} X^T Y) / n$$

Example: Regression with known design distribution

$$\hat{R}_n(\beta) := -\beta^T X^T Y/n + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E} d(X_i\beta)$$

$$\left(\text{In well-specified case: } \mathbb{E}(Y|X) = \dot{d}(X\beta^0) \right)$$

Then

$$W = (X^T Y - \mathbb{E} X^T Y)/n.$$

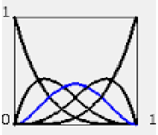
Includes:

- least squares regression
- logistic regression
- Poisson regression

Note: Fixed design is special case of known design distribution

Example: Density estimation

- Let $X_1, \ldots, X_n$ be i.i.d. with distribution P and density $f^0 := dP/d\nu$.
- Let $\{\psi_j(\cdot)\}_{j=1}^p$ be a given dictionary and $\psi(\cdot) := (\psi_1(\cdot), \ldots, \psi_p(\cdot))$.
- Let $\| \cdot \|_\nu$ be the $L_2(\nu)$-norm.



Take

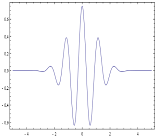
$$\hat{R}_n(\beta) := -\frac{1}{n} \sum_{i=1}^{n} \psi(X_i)\beta + \|\psi\beta\|_\nu^2 / 2$$

Then

$$W = \frac{1}{n} \sum_{i=1}^{n} \psi^T(X_i) - \mathbb{E}\psi^T(X_1).$$

Example: Log-density estimation

- Let $X_1, \ldots, X_n$ be i.i.d. with distribution P and log-density $\log dP/d\nu \propto f^0$
- Let $\{\psi_j(\cdot)\}_{j=1}^p$ be a given dictionary and $\psi(\cdot) := (\psi_1(\cdot), \ldots, \psi_p(\cdot))$.
- Let $d(\psi\beta) := \log[\int \mathrm{e}^{\psi\beta} d\nu]$



Take

$$\hat{R}_n(\beta) := -\frac{1}{n}\sum_{i=1}^{n} \psi(X_i)\beta + d(\psi\beta)$$

Then

$$W = \frac{1}{n}\sum_{i=1}^{n} \psi^T(X_i) - \mathbb{E}\psi^T(X_1).$$

Examples with non-linear “random part”:

- Least absolute deviations
- Probit
- Huber
- mixture models
- PCA
- ...

Examples of norms used

$\boxed{\ell_1\text{-norm:}}$ $\Omega(\beta) = \|\beta\|_1 =: \sum_{j=1}^{p} |\beta_j|$

Examples of norms used

$\boxed{\ell_1\text{-norm:}}$ $\Omega(\beta) = \|\beta\|_1 =: \sum_{j=1}^{p} |\beta_j|$

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

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

[Bondell and Reich 2008]

Examples of norms used

$\boxed{\ell_1\text{-norm:}}$ $\Omega(\beta) = \|\beta\|_1 =: \sum_{j=1}^{p} |\beta_j|$

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

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

[Bondell and Reich 2008]

$\boxed{\text{sorted } \ell_1\text{-norm:}}$ given $\lambda_1 \geq \cdots \geq \lambda_p > 0$,

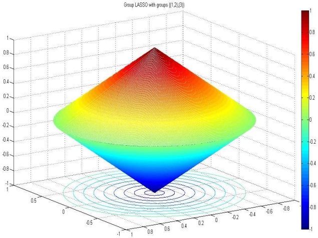
$$\Omega(\beta) := \sum_{j=1}^{p} \lambda_j |\beta|_{(j)} \qquad \text{where } |\beta|_{(1)} \geq \cdots \geq |\beta|_{(p)}$$

[Bogdan et al. 2013]

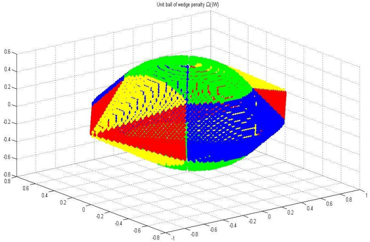
norms generated from cones:

$$\Omega(\beta) := \min_{a \in \mathcal{A}} \frac{1}{2} \sum_{j=1}^{p} \left[\frac{\beta_j^2}{a_j} + a_j \right], \mathcal{A} \subset \mathbb{R}_+^p$$

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



unit ball for group Lasso norm



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

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

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

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

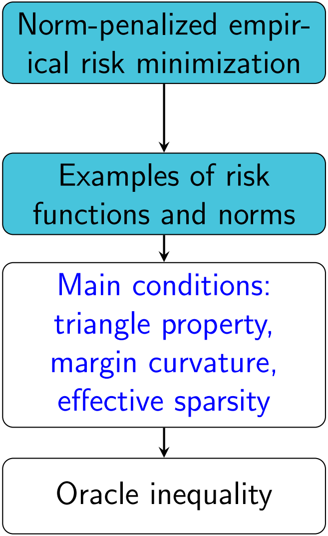
$$\Omega(B) := \|B\|_{\mathrm{nuclear}} := \mathrm{trace}(\sqrt{B^T B})$$

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

$\Omega(B) :=$ dual norm of Ω_*
where

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

[Yuan and Zhang 2014]



Norm-penalized empirical risk minimization

Examples of risk functions and norms

Main conditions: triangle property, margin curvature, effective sparsity

Oracle inequality

Illustations:

Lasso
Matrix completion
Sparse PCA

A sharp oracle inequality: preview

Theorem

Let

$$\overset{\overset{\text{tuning parameter}}{\downarrow}}{\lambda} > \lambda_\epsilon \ge \underline{\Omega}_*(W).$$

Define for some $0 \le \delta < 1$

$$\underline{\lambda} := \lambda - \lambda_\epsilon,\ \bar{\lambda} := \lambda + \lambda_\epsilon + \delta \underline{\lambda},\ L = \frac{\bar{\lambda}}{(1-\delta)\underline{\lambda}}.$$

Then

$$\delta \underline{\lambda}\ \underline{\Omega}(\hat{\beta} - \beta^*) + \mathcal{E}(\hat{\beta}) \le \mathcal{E}(\overset{\overset{\beta^* := \text{ “candidate oracle”}}{\downarrow}}{\beta^*}) + H(\bar{\lambda}\Gamma(L, \beta^*)).$$

A sharp oracle inequality: preview

Theorem

Let

$$\overset{\text{tuning parameter}}{\overset{\downarrow}{\lambda}} > \lambda_\epsilon \geq \underline{\Omega}_*(W).$$

Define for some $0 \leq \delta < 1$

$$\underline{\lambda} := \lambda - \lambda_\epsilon, \ \bar{\lambda} := \lambda + \lambda_\epsilon + \delta\underline{\lambda}, \ L = \frac{\bar{\lambda}}{(1-\delta)\underline{\lambda}}.$$

Then

$$\delta\underline{\lambda}\ \underline{\Omega}(\hat{\beta} - \beta^*) + \mathcal{E}(\hat{\beta}) \leq \mathcal{E}(\overset{\beta^* := \text{ “candidate oracle”}}{\overset{\downarrow}{\beta^*}}) + H(\bar{\lambda}\Gamma(L, \beta^*)).$$

W= “random part”

$\underline{\Omega}_*$ dual of $\underline{\Omega} = \Omega^+ + \Omega^-$

$\Gamma(L, \beta^*)$=effective sparsity

H=convex conjugate of margin curvature


triangle property

everything to explain

Two point inequality and triangle property

Let $\dot{\hat{R}}_n(\beta) := \frac{\partial}{\partial \beta} \hat{R}_n(\beta)$

Lemma *We have the two point inequality*

$$-\dot{\hat{R}}_n(\hat{\beta})^T(\beta - \hat{\beta}) \leq \lambda \quad \underbrace{\left(\Omega(\beta) - \Omega(\hat{\beta}) \right)}_{\textit{this we need to control}} \quad \forall \ \beta.$$

Control of the term $\Bigg(\underset{\substack{\uparrow\\ \text{fixed } \beta^+}}{\Omega(\beta)} - \underset{\substack{\uparrow\\ \text{variable } \beta}}{\Omega(\hat{\beta})}\Bigg)$ [1]

Triangle inequality:

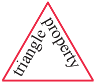
$$\Bigg(\Omega(\beta^+) - \Omega(\beta)\Bigg) \leq \Omega(\beta^+ - \beta)$$

... is typically to rough

Definition

We say that the triangle property holds at β^+
if for some semi-norms Ω^+ *and* Ω^-

$$\Bigg(\Omega(\beta^+) - \Omega(\beta)\Bigg) \leq \Omega^+(\beta^+ - \beta) - \Omega^-(\beta), \forall\ \beta.$$



---

[1] β^+ may be a candidate oracle β^*

Examples of the triangle property

Notation

For a set $S \subset \{1, \ldots, p\}$

$$\beta := \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{j-1} \\ \beta_j \\ \beta_{j+1} \\ \vdots \\ \beta_p \end{pmatrix} \quad \beta_S := \begin{pmatrix} \beta_1 \\ \vdots \\ 0 \\ 0 \\ \beta_{j+1} \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} \leftarrow \in S \\ \vdots \\ \leftarrow \notin S \\ \leftarrow \notin S \\ \leftarrow \in S \\ \vdots \\ \leftarrow \notin S \end{matrix}$$

Examples of the triangle property

Notation

For a set $S \subset \{1, \ldots, p\}$

$$\beta := \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{j-1} \\ \beta_j \\ \beta_{j+1} \\ \vdots \\ \beta_p \end{pmatrix} \quad \beta_S := \begin{pmatrix} \beta_1 \\ \vdots \\ 0 \\ 0 \\ \beta_{j+1} \\ \vdots \\ 0 \end{pmatrix} \quad \begin{matrix} \leftarrow \in S \rightarrow \\ \vdots \\ \leftarrow \notin S \rightarrow \\ \leftarrow \notin S \rightarrow \\ \leftarrow \in S \rightarrow \\ \vdots \\ \leftarrow \notin S \rightarrow \end{matrix} \quad \begin{pmatrix} 0 \\ \vdots \\ \beta_{j-1} \\ \beta_j \\ 0 \\ \vdots \\ \beta_p \end{pmatrix} =: \beta_{-S}$$

ℓ_1-norm

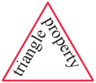
Let $S := \{j : \ \beta_j^+ \ne 0\}$ ($=$ active set of β^+).

Triangle property:

$$\|\beta^+\|_1 - \|\beta\|_1 \le \underbrace{\sum_{j \in S} |\beta_j^+ - \beta_j|}_{\Omega^+(\beta^+ - \beta)} - \underbrace{\sum_{j \notin S} |\beta_j|}_{\Omega^-(\beta)}$$

or

$$\|\beta^+\|_1 - \|\beta\|_1 \le \|\beta_S^+ - \beta_S\|_1 - \|\beta_{-S}\|_1$$



sorted ℓ_1-norm / Oscar:

$\Omega(\beta) := \sum_{j=1}^{p} \lambda_j |\beta|_{(j)}$.
Let $S := \{j : \ \beta^+_j \neq 0\}$ and $s := |S|$.
Triangle property:

$$\Omega(\beta^+) - \Omega(\beta) \le \underbrace{\Omega(\beta^+_S - \beta_S)}_{\Omega^+(\beta^+ - \beta)} - \underbrace{\sum_{j=1}^{p-s} \lambda_{s+j} |\beta|_{(j,-S)}}_{\Omega^-(\beta)}$$

norms generated from cones:

$$\Omega(\beta) := \min_{a\in\mathcal{A}} \frac{1}{2}\sum_{j=1}^{p}\left[\frac{\beta_j^2}{a_j} + a_j\right].$$

Let $S := \{j : \ \beta_j^+ \neq 0\}$.
Suppose $\{a_S : \ a \in \mathcal{A}\} \subset \mathcal{A}$.

Triangle property:

$$\Omega(\beta^+) - \Omega(\beta) \le \underbrace{\Omega(\beta_S^+ - \beta_S)}_{\Omega^+(\beta^+ - \beta)} - \underbrace{\min_{a_{-S}\in\mathcal{A}_{-S}} \frac{1}{2}\sum_{j\notin S}\left[\frac{\beta_j^2}{a_j} + a_j\right]}_{\Omega^-(\beta)}$$

nuclear norm for matrices:

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

Let B^+ have SVD

$$B^+ = U\Phi V^T, \; U^TU = I, \; V^TV = I, \; \Phi = \text{diag}(\phi_1, \ldots, \phi_s)$$

<u>Triangle property:</u> for $P_1 := UU^T$, $P_2 := VV^T$

$$\begin{aligned} \Omega(B^+) - \Omega(B) \;\; &\leq \;\; \underbrace{\|P_1(B^+ - B)P_2\|_{\text{nuclear}}}_{(\leq)\Omega^+(B^+ - B)} \\ &- \;\; \underbrace{\|P_1^\perp B P_2^\perp\|_{\text{nuclear}}}_{\Omega^-(B)} \end{aligned}$$

nuclear norm for tensors:

Triangle property:

$$\Omega(B^+) - \Omega(B) \le \underbrace{\Omega(\mathcal{Q}^0(B^+ - B))}_{(\le)\Omega^+(B^+ - B)} - \underbrace{\tfrac{1}{2}\,\Omega(\mathcal{Q}^\perp B)}_{\Omega^-(B)},$$

$\mathcal{Q}_0 := P_1 \otimes P_1 \otimes P_3,$
$\mathcal{Q}_1^\perp := P_1 \otimes P_2^\perp \otimes P_3^\perp$
$\mathcal{Q}_2^\perp := P_1^\perp \otimes P_2 \otimes P_3^\perp$
$\mathcal{Q}_3^\perp := P_1^\perp \otimes P_2^\perp \otimes P_3$
$\mathcal{Q}^\perp := \mathcal{Q}_0^\perp + \mathcal{Q}_1^\perp + \mathcal{Q}_2^\perp + \mathcal{Q}_3^\perp$
[Yuan and Zhang, 2014]

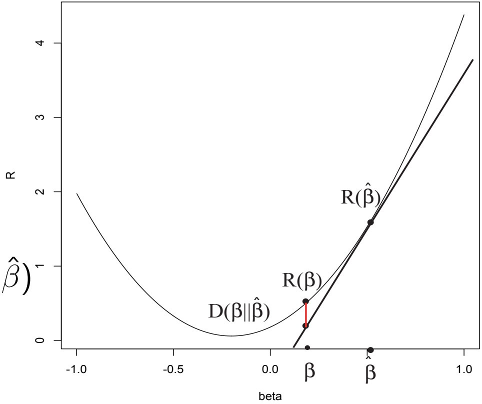
Margin curvature and effective sparsity

Let $\dot{R}(\beta) := \frac{\partial}{\partial \beta} R(\beta)$

The Bregman divergence is

$$D(\beta \| \hat{\beta})$$

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



Bregman divergence $D(\beta \| \hat{\beta}) = R(\beta) - R(\hat{\beta}) - \dot{R}(\hat{\beta})^T(\beta - \hat{\beta})$

Two point margin condition (at some β^*)

There is a semi-norm τ *on* $\mathbb{R}^p$
and a strictly convex function G $\nearrow$
such that

$$D(\beta^* \| \hat{\beta}) \geq G(\tau(\beta^* - \hat{\beta}))$$

We call τ the *margin semi-norm* and G the *margin curvature*.

Definition *The effective sparsity at* β^+ *is*

$$\Gamma^2(L, \beta^+) := \max\left\{ \left(\frac{\Omega^+(\beta)}{\tau(\beta)} \right)^2 : \underbrace{\Omega^-(\beta) \leq L\Omega^+(\beta)}_{\text{“cone condition”}} \right\}.$$

Here $L \geq 1$ *is a stretching factor.*

We call this an Ω / τ comparison (at β^+)

$\boxed{\text{Wald lecture 3 (Friday):}}$

Discussion of bounds for $\Gamma^2(L, \beta^+)$

Example:

ℓ_1 / ℓ_2 comparison

- $\Omega = \|\cdot\|_1$ and $\tau(\beta) := \|\beta\|_2$
- $S := \{j : \ \beta_j^+ \neq 0\}$
- $s := |S|$
- $\Omega^+(\beta) := \sum_{j\in S} |\beta_j|$
- $\Omega^-(\beta) := \sum_{j\notin S} |\beta_j|$

$\underset{\text{Lemma}}{\Rightarrow} \boxed{\Gamma^2(L, \beta^+) = s}$:

Example:

ℓ_1 / ℓ_2 comparison

- ○ $\Omega = \|\cdot\|_1$ and $\tau(\beta) := \|\beta\|_2$
- ○ $S := \{j : \ \beta_j^+ \neq 0\}$
- ○ $s := |S|$
- ○ $\Omega^+(\beta) := \sum_{j \in S} |\beta_j|$
- ○ $\Omega^-(\beta) := \sum_{j \notin S} |\beta_j|$

$\underset{\text{Lemma}}{\Rightarrow}$ $\boxed{\Gamma^2(L, \beta^+) = s}$:

Proof.

$$\begin{aligned} \|\beta_S\|_1^2 &= \left(\sum_{j \in S} |\beta_j| \right)^2 \\ &\leq s \sum_{j \in S} |\beta_j|^2 \\ &\leq s \sum_{j=1}^{p} |\beta_j|^2 = s\|\beta\|_2^2 \end{aligned}$$

$$\boxed{\Rightarrow \max\{\|\beta_S\|_1^2 / \|\beta\|_2^2 : \ \|\beta_{-S}\|_1 \leq L\|\beta_S\|_1\} = s}$$

□

we have now defined

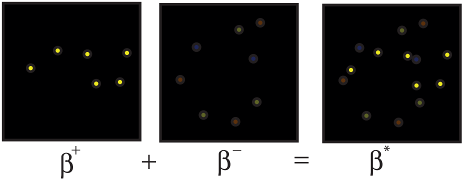
- the triangle property
- the margin curvature
- the effective sparsity

still to define

- the candidate oracle

The candidate oracle β^*

- Fix β^+ and let $\underline{\Omega} := \Omega^+ + \Omega^-$.
- Let $\Omega^+(\beta^-) = 0$.
- A candidate oracle is then $\beta^+ + \beta^- =: \beta^*$.



- Let $\underline{\Omega}_*$ be the dual norm of $\underline{\Omega}$.

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

$$\beta^* = (\underbrace{*, \cdots, \overset{\overset{s}{\downarrow}}{*}}_{\beta^+}, \underbrace{* \cdots, *}_{\beta^-}) \qquad \text{(transposed)}$$

- $\beta^* \in \mathbb{R}^p$ arbitrary
- $s \in \{1, \ldots, p\}$ arbitrary
- $S :=$ indices of largest $|\beta_j^*|$
- $\Omega^+(\beta) := \|\beta_S\|_1$
- $\Omega^-(\beta) := \|\beta_{-S}\|_1$
- $\beta^+ := \beta_S^*$
- $\beta^- := \beta_{-S}^*$

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

$$\beta^* = (\underbrace{*, \cdots, \overset{\overset{s}{\downarrow}}{*}}_{\beta^+}, \underbrace{* \cdots, *}_{\beta^-}) \qquad \text{(transposed)}$$

- $\beta^* \in \mathbb{R}^p$ arbitrary
- $s \in \{1, \ldots, p\}$ arbitrary
- $S :=$ indices of largest $|\beta_j^*|$
- $\Omega^+(\beta) := \|\beta_S\|_1$
- $\Omega^-(\beta) := \|\beta_{-S}\|_1$
- $\beta^+ := \beta_S^*$
- $\beta^- := \beta_{-S}^*$

Then

- $\underline{\Omega} = \Omega = \| \cdot \|_1$
- $\Omega^+(\beta^-) = \|(\beta_{-S}^*)_S\|_1 = 0$
- $\beta^* = \beta_S^* + \beta_{-S}^* = \beta^+ + \beta^-$
- $\underline{\Omega}_* = \| \cdot \|_\infty$

Example

$\Omega = \| \cdot \|_{\mathrm{nuclear}}, \; p_1 \geq p_2$

$$
\begin{aligned}
B^* &= U \begin{pmatrix} * & & & & \\ & * & & & \\ & & * & & \\ & & & \ddots & \\ & & & & * \end{pmatrix} V^T \\
&= \underbrace{U \begin{pmatrix} * & & & & \\ & * & & & \\ & & 0 & & \\ & & & \ddots & \\ & & & & 0 \end{pmatrix} V^T}_{B^+} + \underbrace{U \begin{pmatrix} 0 & & & & \\ & 0 & & & \\ & & * & & \\ & & & \ddots & \\ & & & & * \end{pmatrix} V^T}_{B^-}
\end{aligned}
$$

Example $\Omega = \|\cdot\|_{\text{nuclear}}$, $p_1 \geq p_2$

- $B^* \in \mathbb{R}^{p_1 \times p_2}$ arbitrary
- $B^* := U\Phi^* V^T$ SVD
- $s \in \{1, \ldots, p_2\}$ arbitrary
- $S := \{1, \ldots, s\}$
- $P_1 := U_S U_S^T$, $P_2 := V_S V_S^T$
- $\Omega^+(B) \geq \|P_1 B P_2\|_{\text{nuclear}}$
- $\Omega^-(B) := \|P_1^\perp B P_2^\perp\|_{\text{nuclear}}$
- $B^+ := U\Phi_S^* V^T$
- $B^- := U\Phi_{-S}^* V^T$

- $\Lambda_{\max}(\cdot) :=$ largest principal component

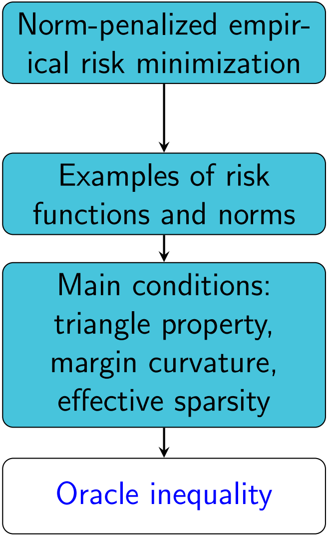
Example $\Omega = \|\cdot\|_{\text{nuclear}}$, $p_1 \geq p_2$

- $B^* \in \mathbb{R}^{p_1 \times p_2}$ arbitrary
- $B^* := U\Phi^* V^T$ SVD
- $s \in \{1, \ldots, p_2\}$ arbitrary
- $S := \{1, \ldots, s\}$
- $P_1 := U_S U_S^T$, $P_2 := V_S V_S^T$
- $\Omega^+(B) \geq \|P_1 B P_2\|_{\text{nuclear}}$
- $\Omega^-(B) := \|P_1^\perp B P_2^\perp\|_{\text{nuclear}}$
- $B^+ := U\Phi_S^* V^T$
- $B^- := U\Phi_{-S}^* V^T$

Then

- $\underline{\Omega} \leq \Omega = \|\cdot\|_{\text{nuclear}}$
- $\Omega^+(B^-) = 0$
- $B^* = B^+ + B^-$
- $\underline{\Omega}_* \geq \Omega_* = \Lambda_{\max}$

- $\boxed{\Lambda_{\max}(\cdot) := \text{largest principal component}}$



Norm-penalized empirical risk minimization

Examples of risk functions and norms

Main conditions: triangle property, margin curvature, effective sparsity

Oracle inequality

Illustations:

Lasso

Matrix completion

Sparse PCA

Recap

We say that the "random part" $\hat{R}_n(\beta) - R(\beta)$ *is linear if for a random vector* $\begin{pmatrix} W_0 \\ W \end{pmatrix} \in \mathbb{R}^{p+1}$

$$R_n(\beta) - R(\beta) = W_0 - \beta^T W \ \ \forall \ \beta$$

.

Further:

- $\underline{\Omega} := \Omega^+ + \Omega^-$

- $\underline{\Omega}_*$: dual norm
- G: margin curvature
- $\Gamma^2(L, \beta^+)$: effective sparsity

A sharp oracle inequality

Theorem [vdG, 2016] *Suppose the "random part" is linear. Let*

$$\lambda > \lambda_\epsilon \geq \underline{\Omega}_*(W).$$

Define for some $0 \leq \delta < 1$

$$\underline{\lambda} := \lambda - \lambda_\epsilon, \ \bar{\lambda} := \lambda + \lambda_\epsilon + \delta \underline{\lambda}, \ L = \frac{\bar{\lambda}}{(1-\delta)\underline{\lambda}}.$$

Then[2]

$$\delta \underline{\lambda \Omega}(\hat{\beta} - \beta^*) + R(\hat{\beta}) \leq R(\beta^*) + H(\bar{\lambda}\Gamma(L, \beta^+)) + 2\lambda\Omega(\beta^-)$$

where H is the convex conjugate of G.

---

[2] Recall: the excess risk is $\mathcal{E}(\beta) = R(\beta) - R(\beta^0)$

A sharp oracle inequality

Theorem (Extension to possibly nonlinear "random part")
Let

$$\lambda > \lambda_\epsilon \geq \underline{\Omega}_*(\dot{\hat{R}}_n(\hat{\beta}) - \dot{R}(\hat{\beta}))..$$

Define for some $0 \leq \delta < 1$

$$\underline{\lambda} := \lambda - \lambda_\epsilon,\ \ \bar{\lambda} := \lambda + \lambda_\epsilon + \delta\underline{\lambda},\ \ L = \frac{\bar{\lambda}}{(1-\delta)\underline{\lambda}}.$$

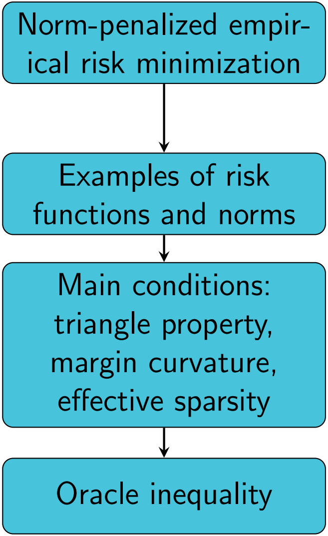
Then[3]

$$\delta\underline{\lambda\Omega}(\hat{\beta} - \beta^*) + R(\hat{\beta}) \leq R(\beta^*) + H(\bar{\lambda}\Gamma(L, \beta^+)) + 2\lambda\Omega(\beta^-)$$

where H *is the convex conjugate of* G.

---

[3] Recall: the excess risk is $\mathcal{E}(\beta) = R(\beta) - R(\beta^0)$



Norm-penalized empirical risk minimization

Examples of risk functions and norms

Main conditions: triangle property, margin curvature, effective sparsity

Oracle inequality

Illustations:

Lasso

Matrix completion

Sparse PCA

Example: Lasso

Model: $Y = X\beta^0 + \epsilon$, fixed design

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \Big\{ \underbrace{\|Y - X\beta\|_2^2/(2n)}_{\hat{R}_n(\beta)} + \lambda \underbrace{\|\beta\|_1}_{\Omega(\beta)} \Big\}$$

Dual norm: $\lambda_\epsilon := \sqrt{2\log(2p/\alpha)/n} \overset{\text{with probability } \geq 1-\alpha}{\overset{\downarrow}{\geq}} \|W\|_\infty, \ W = X^T\epsilon/n$

Margin semi-norm: $\tau^2(\beta) = \|X\beta\|_2^2/n$

Margin curvature: $G(u) = u^2/2 \Rightarrow H(v) = v^2/2$

Effective sparsity: $\Gamma^2(L, \beta^+) = s/\hat{\phi}^2(L, S)$

where

$$\begin{aligned} \hat{\phi}^2(L,S) \quad &:= \quad \min\{s\|X\beta\|_2^2/n : \ \|\beta_S\|_1 = 1, \ \|\beta_{-S}\|_1 \leq L\} \\ &= \quad \text{“compatibility constant”} \end{aligned}$$

From the theorem: with probability $\geq 1 - \alpha$

$$\delta\underline{\lambda}\|\hat{\beta} - \beta^*\|_1 + \frac{\|X(\hat{\beta} - \beta^0)\|_2^2}{2n}$$

$$\leq \underbrace{\frac{\|X(\beta^* - \beta^0)\|_2^2}{2n}}_{\text{approximation error}} + \underbrace{\frac{s\bar{\lambda}^2}{2\hat{\phi}^2(L, S)}}_{\text{estimation error}} + \underbrace{2\lambda\|\beta^*_{-S}\|_1}_{\text{smallish coefficients}}$$

Taking $\beta^* = \beta^0$ and minimizing over S gives

$$\delta\underline{\lambda}\|\hat{\beta} - \beta^0\|_1 + \frac{\|X(\hat{\beta} - \beta^0)\|_2^2}{2n}$$

$$\leq \min_{S\subset\{1,\ldots,p\}}\left\{\frac{s\bar{\lambda}^2}{2\hat{\phi}^2(L, S)} + 2\lambda\|\beta^0_{-S}\|_1\right\}$$

From the theorem: with probability $\geq 1-\alpha$

$$\delta\underline{\lambda}\|\hat{\beta}-\beta^*\|_1 + \frac{\|X(\hat{\beta}-\beta^0)\|_2^2}{2n}$$

$$\leq \underbrace{\frac{\|X(\beta^*-\beta^0)\|_2^2}{2n}}_{\text{approximation error}} + \underbrace{\frac{s\bar{\lambda}^2}{2\hat{\phi}^2(L,S)}}_{\text{estimation error}} + \underbrace{2\lambda\|\beta^*_{-S}\|_1}_{\text{smallish coefficients}}$$

Taking $\beta^* = \beta^0$ and minimizing over S gives

$$\delta\underline{\lambda}\|\hat{\beta}-\beta^0\|_1 + \frac{\|X(\hat{\beta}-\beta^0)\|_2^2}{2n}$$

$$\leq \min_{S\subset\{1,\ldots,p\}} \left\{ \frac{s\bar{\lambda}^2}{2\hat{\phi}^2(L,S)} + 2\lambda\|\beta^0_{-S}\|_1 \right\}$$

$$\underset{\text{by taking } S=S_0}{\leq} \frac{s_0\bar{\lambda}^2}{2\hat{\phi}^2(L,S_0)}$$

Comparison Lasso and sorted ℓ_1-norm (Slope)

[4]

Table: Fixed β

	theoretical λ			Cross-validated λ		
	$\|\beta^0 - \hat{\beta}\|_{\ell_1}$	$\Omega(\beta^0 - \hat{\beta})$	$\|X(\beta^0 - \hat{\beta})\|_{\ell_2}$	$\|\beta^0 - \hat{\beta}\|_{\ell_1}$	$\Omega(\beta^0 - \hat{\beta})$	$\|X(\beta^0 - \hat{\beta})\|_{\ell_2}$
srSLOPE	2.06	0.21	4.12	2.37	0.26	3.88
srLASSO	1.85	0.19	5.51	1.78	0.19	5.05

Table: Random β

	theoretical λ			cross-validated λ		
	$\|\beta^0 - \hat{\beta}\|_1$	$\Omega(\beta^0 - \hat{\beta})$	$\|X(\beta^0 - \hat{\beta})\|_{\ell_2}$	$\|\beta^0 - \hat{\beta}\|_1$	$\Omega(\beta^0 - \hat{\beta})$	$\|X(\beta^0 - \hat{\beta})\|_{\ell_2}$
srSLOPE	4.50	0.49	7.74	7.87	1.09	7.68
srLASSO	8.48	0.89	29.47	7.81	0.85	9.19

---

[4] [Stucky and vdG, 2016]

Norm-penalized empirical risk minimization

Examples of risk functions and norms

Main conditions: triangle property, margin curvature, effective sparsity

Oracle inequality

Illustations:

- Lasso
- Matrix completion
- Sparse PCA

Example: Matrix completion in logistic regression [Lafond, 2015]

Let X_i be a mask with a "1" at a random entry.

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

$$\hat{R}_n(B) := -\frac{1}{n} \sum_{i=1}^{n} Y_i \mathrm{trace}(X_i B) + \sum_{j,k} d(B_{j,k})/(p_1 p_2),$$

where

- $B \in \mathcal{B} := \{ B \in \mathbb{R}^{p_1 \times p_2} : \ \eta \le \| B \|_\infty \le 1 - \eta \}$, $0 < \overset{\text{given}}{\overset{\downarrow}{\eta}} < 1$
- $d(\xi) := \log(1 + \mathrm{e}^{\xi}), \ \xi \in \mathbb{R}$

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

Dual norm: $\lambda_\epsilon \geq \Lambda_{\max}(W), \; W = \sum_{i=1}^n (X_i^T Y_i - \mathbb{E}(X_i^T Y_i))/n$

Margin semi-norm: $\tau^2(B) = \|B\|_2^2/(p_1 p_2)$

Margin curvature: $G(u) = u^2/(2c p_1 p_2) \Rightarrow H(v) = c p_1 p_2 v^2/2$

Effective sparsity: $\Gamma^2(L, \beta^+) = 3s$

From the theorem:

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 $1 - \alpha$

$$\delta \underline{\lambda} \| \hat{B} - B^* \|_{\text{nuclear}} + R(\hat{B}) \leq R(B^*) + C \times \overset{\substack{\text{estimation}\\ \text{error}\\ \downarrow}}{\left(\frac{p_1 \log(p_1) s}{n} \right)} + \overset{\substack{\text{smallish}\\ \text{singular values}\\ \downarrow}}{2\lambda \| \phi^*_{-S} \|_1}.$$

Taking $B^* = B^0$ gives

$$\begin{aligned} &\delta \underline{\lambda} \| \hat{B} - B^* \|_{\text{nuclear}} + R(\hat{B}) - R(B^0) \\ &\leq \min_{S \subset \{1, \ldots, p_2\}} \left\{ C \times \left(\frac{p_1 \log(p_1) s}{n} \right) + 2\lambda \| \phi^0_{-S} \|_1 \right\} \\ &\leq C \times \left(\frac{p_1 \log(p_1) s_0}{n} \right) \end{aligned}$$

In the previous example we assumed that the distribution of the design is known.
And we had a linear “random part”.

In the next example we no longer assume the distribution of the design to be known.
And we have a non-linear “random part”.

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

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

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

where

- $B \in \mathcal{B} := \{B \in \mathbb{R}^{p_1 \times p_2} : \ \|B\|_\infty \leq \eta\}$ 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:
$G(u) = u^2/(2 c p_1 p_2)$
Effective sparsity:
$\Gamma^2(L, \beta^+) = 3s$

as before

From the theorem:

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 $1 - \alpha$

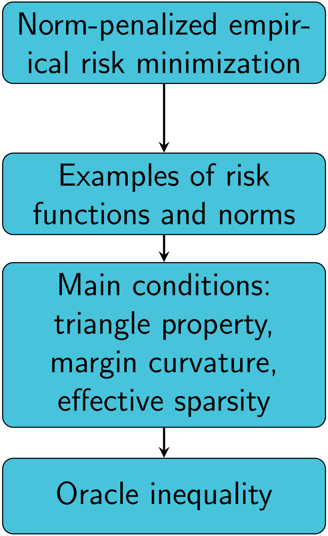
$$\delta\underline{\lambda}\|\hat{B} - B^*\|_{\mathrm{nuclear}} + R(\hat{B}) \leq R(B^*) + C \times \left(\frac{p_1 \log(p_1) s}{n}\right) + 2\lambda\|\phi^*_{-S}\|_1.$$

everything as before

BUT: the estimator does not require knowing the distribution of the design.
If the masks X_i are not uniformly distributed we get a different normalization.

Huber loss is twice differentiable
Least absolute deviations loss is not

$\leadsto$ $\underline{\text{nonsharp}}$ oracle inequality
for matrix completion with least absolute deviations loss
[Elsener and vdG (2016)]



Norm-penalized empirical risk minimization

Examples of risk functions and norms

Main conditions: triangle property, margin curvature, effective sparsity

Oracle inequality

Illustations:

- Lasso
- Matrix completion
- Sparse PCA

Example: Sparse PCA

- $X_1, \dots, X_n$ i.i.d. $\in \mathbb{R}^p$
- $\hat{\Sigma} := X^T X / n$
- $\hat{R}_n(\beta) := \|\hat{\Sigma} - \beta\beta^T\|_2^2$
- $\Omega := \|\cdot\|_1$

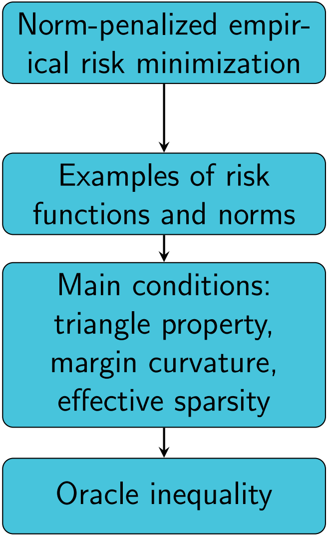
From the theorem:

Assume

- ○ $\mathcal{B} \subset \{\|\beta - \beta^0\|_2 \le \eta\}$
- ○ spikiness
- ○ $p < n$ or $\mathcal{B} \subset \{\|\beta - \beta^0\|_1 \le \eta\}$
- ○ e.g. bounded design

Then for $s = |S| = o(\sqrt{n/\log p})$, $\lambda = C_0\sqrt{\log p/n}$, w.h.p.

$$\delta\underline{\lambda}\|\beta^* - \beta\|_1 + R(\hat{\beta}) \le R(\beta^*) + \bar{\lambda}^2 s_*/8 + 2\lambda\|\beta^*_{-S}\|_1$$



Norm-penalized empirical risk minimization

Examples of risk functions and norms

Main conditions: triangle property, margin curvature, effective sparsity

Oracle inequality

Illustations:

Lasso

Matrix completion

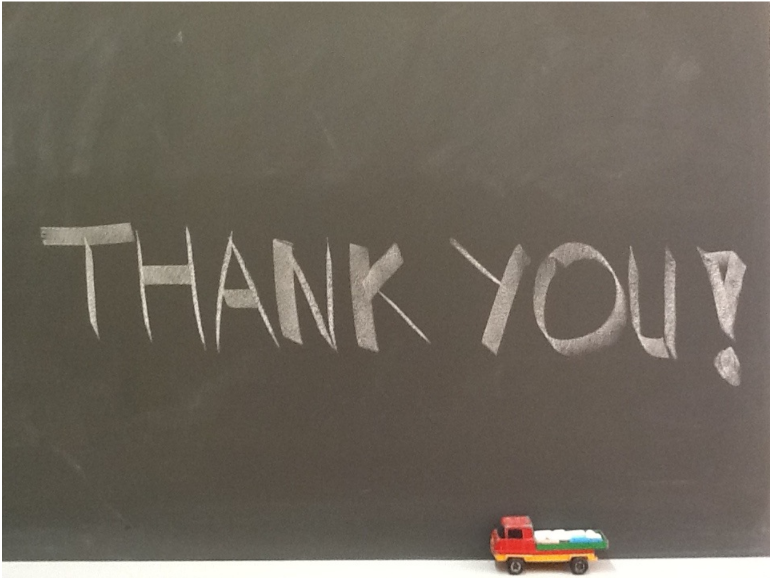
Sparse PCA

Conclusion

norms with the triangle property

lead to oracle inequalities

for general loss and assuming margin curvature


THANK YOU!