Some bias and a pinch of variance

Sara van de Geer

November 2, 2016

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

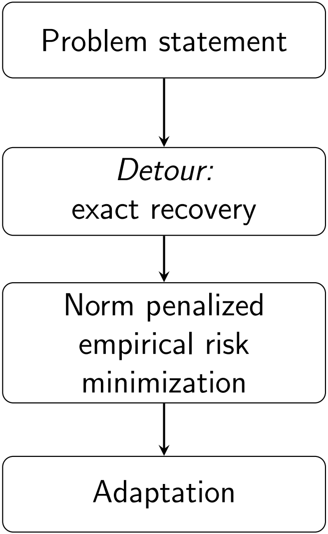
... this talk is about theory for machine learning algorithms ...

... this talk is about theory for machine learning algorithms ...
... for high-dimensional data ...

... it is about prediction performance of algorithms trained on random data ...

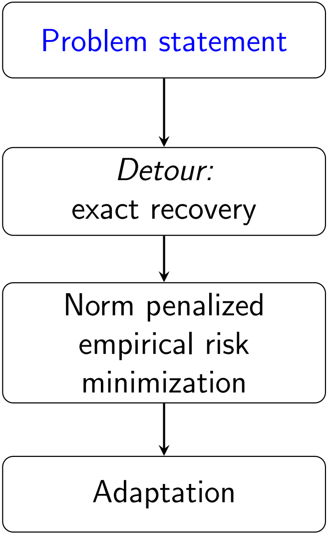
it is
not about
the
scripts
used

```
procedure Transpose (a)Order:(n) ; value n ;
array a ; integer n ;
begin real w ; integer i, k ;
for i := 1 step 1 until n do
    for k := 1+i step 1 until n do
    begin w := a[i,k] ;
          a[i,k] := a[k,i] ;
          a[k,i] := w
    end
end Transpose
```



Concepts:

Sparsity
Effective sparsity
Margin Curvature
Triangle property



Concepts:

Sparsity
Effective sparsity
Margin Curvature
Triangle property

Problem:

Let

$$f:\ \mathcal{X} \to \mathbb{R},\ \mathcal{X} \subset \mathbb{R}^m$$

Find

$$\min_{x \in \mathcal{X}} f(x)$$

Problem:

Let

$$f: \mathcal{X} \to \mathbb{R}, \; \mathcal{X} \subset \mathbb{R}^m$$

Find

$$\min_{x \in \mathcal{X}} f(x)$$

Severe Problem:

The function f is unknown!

What we <u>do</u> know:

$$f(x) = \int \ell(x,y)dP(y) =: f_P(x)$$

where

∘ $\ell(x,y)$ is a given "loss" function:

$$\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

∘ P is an unknown probability measure on the space $\mathcal{Y}$

Example

- $\mathcal{X} :=$ the persons you consider marrying
- $\mathcal{Y} :=$ possible states of the world
- $\ell(x, y) :=$ the loss when marrying x in world y
- $P :=$ the distribution of possible states of the world
- $f(x) = \int \ell(x, y) dP(y)$ the "risk" of marrying x

Let Q be a given probability measure on $\mathcal{Y}$

We replace P by Q:

$$f_Q(x) := \int \ell(x, y) dQ(y)$$

and estimate

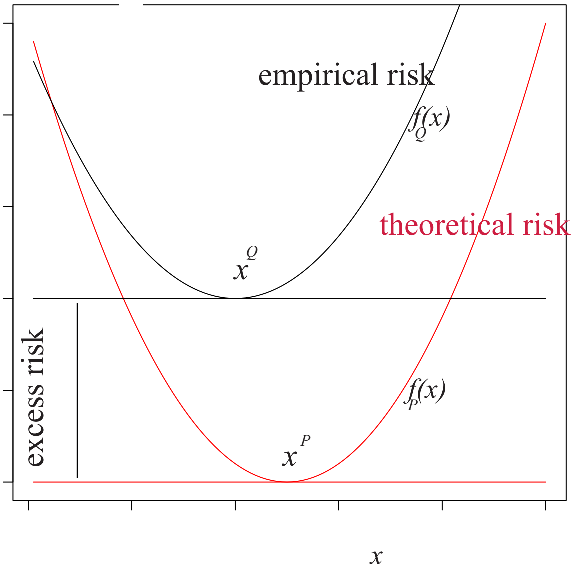
$$x^P := \arg\min_{x \in \mathcal{X}} f_P(x)$$

by

$$x^Q := \arg\min_{x \in \mathcal{X}} f_Q(x)$$

Question:

How "good" is this estimate?



empirical risk
f_Q(x)
theoretical risk
x^Q
excess risk
f_P(x)
x^P
x

Question:

Is

x^Q close to x^P ?

$f(x^Q)$ close to $f(x^P)$

... in our setup ...

we have to regularize: accept some bias to reduce variance

Our setup:
$Q :=$ corresponds to a sample $Y_1, \ldots, Y_n$ from P
$n :=$ sample size
Thus

$$f^Q(x) := \hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \ell(x, Y_i), \; x \in \mathcal{X} \subset \mathbb{R}^m$$

(a *random* function)

number of parameters

$$m$$

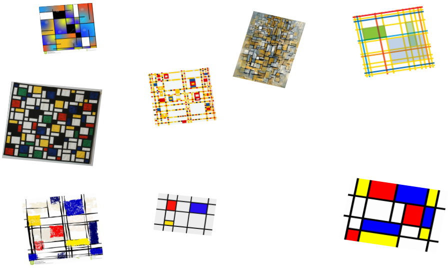
number of observations

$$n$$

high-dimensional statistics:

$$m \gg n$$

DATA



$$Y_1, \ldots, Y_n$$

$$\downarrow$$

$$\hat{x} \in \mathbb{R}^m$$

In our setup with $m \gg n$ we need to regularize

That is: accept some bias to be able to reduce the variance.

Regularized empirical risk minimization

Target:

$$x^P := x^0 = \arg \min_{x \in \mathcal{X} \subset \mathbb{R}^m} \underbrace{f_P(x)}_{\text{unobservable risk}}$$

Estimator based on sample:

$$x^Q := \hat{x} := \arg \min_{x \in \mathcal{X} \subset \mathbb{R}^m} \Big\{ \underbrace{f_Q(x)}_{\text{empirical risk}} + \underbrace{\mathrm{pen}(x)}_{\text{regularization penalty}} \Big\}$$

Example:

Let $Z \in \mathbb{R}^{n \times m}$ be a given design matrix
and $b^0 \in \mathbb{R}^n$ unobserved vector

Let $\|v\|_2^2 := \sum_{i=1}^n v_i^2$ and

$$x^0 \in \arg\min_{x \in \mathbb{R}^m} \overbrace{\|b^0 - Zx\|_2^2}^{f_P(x)}$$

Sample

$$Y = b^0 + \epsilon, \ \ \epsilon \in \mathbb{R}^n \text{ noise}$$

"Lasso" with "tuning parameter" $\lambda \geq 0$:

$$\hat{x} := \arg\min_{x \in \mathbb{R}^p} \left\{ \overbrace{\|Y - Zx\|_2^2}^{f_Q(x)} + 2\lambda \overbrace{\|x\|_1}^{:=\sum_{j=1}^m |x_j|} \right\}$$

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

High-dimensional: $m \gg n$

Definition

We call j an active parameter if (roughly speaking) $x_j^0 \neq 0$

We say x^0 is sparse if the number of active parameters is small

We write the active set of x^0 as

$$S_0 := \{j : \; x_j^0 \neq 0\}$$

We call $s_0 := |S_0|$ the sparsity of x^0

Goal:

- derive oracle inequalities
  for norm-penalized empirical risk minimizers
  oracle: an estimator that knows the "true" sparsity
  oracle inequalities:

  Adaptation

  to unknown sparsity

Benchmark

Low-dimensional

$$\hat{x} = \arg \min_{x \in \mathcal{X} \subset \mathbb{R}^m} \hat{f}_n(x)$$

Then typically

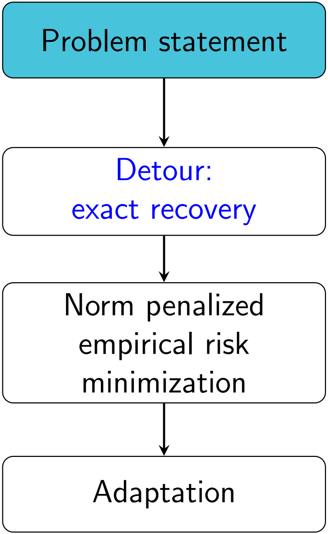
$$f_P(\hat{x}) - f_P(x^0) \sim \frac{m}{n} = \frac{\text{number of parameters}}{\text{number of observations}}$$

High-dimensional

$$\hat{x} = \arg \min_{x \in \mathcal{X} \subset \mathbb{R}^m} \left\{ \hat{f}_n(x) + \text{pen}(x) \right\}$$

Aim is *Adaptation*

$$f_P(\hat{x}) - f_P(x^0) \sim \frac{s_0}{n} = \frac{\text{number of active parameters}}{\text{number of observations}}$$



Concepts:

Sparsity
Effective sparsity
Margin curvature
Triangle property

Exact recovery

Let $Z \in \mathbb{R}^{n \times m}$ be given and $b^0 \in \mathbb{R}^n$ be given with $m \gg n$

Consider the system

$$Zx^0 = b^0$$

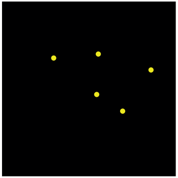
of n equations with m unknowns
Basis pursuit:

$$x^* := \arg\min_{x \in \mathbb{R}^m} \left\{ \|x\|_1 : \ Zx = b^0 \right\}$$

Notation

Active set:

$S_0 := \{j : \ x_j^0 \neq 0\}$



Sparsity:

$s_0 := |S_0|$

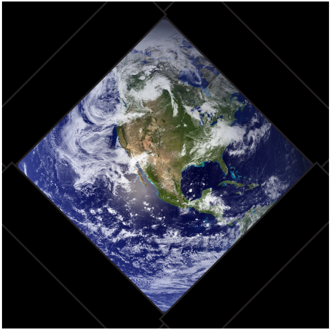
Effective sparsity:

$$\Gamma_0^2 := \frac{s_0}{\hat{\phi}^2(S_0)} = \max\left\{ \frac{\|x_{S_0}\|_1^2}{\|Zx\|_2^2/n} : \ \underbrace{\|x_{-S_0}\|_1 \leq \|x_{S_0}\|_1}_{\text{"cone condition"}} \right\}$$

Compatibility constant: $\hat{\phi}^2(S_0)$

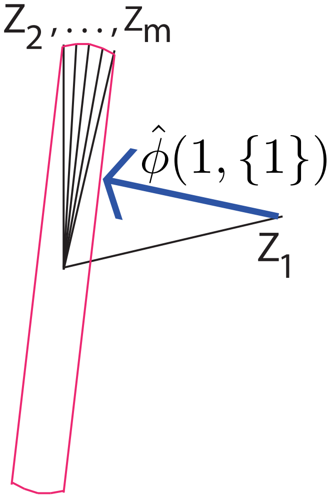
The compatibility constant is *canonical correlation* ...

... in the ℓ_1-world



The effective sparsity Γ_0^2 is $\approx$ the sparsity s_0 but taking into account the correlation between variables.

Compatibility constant: (in $\mathbb{R}^2$)



$\hat{\phi}(S) = \hat{\phi}(1, S)$ for the case $S = \{1\}$

Basis Pursuit

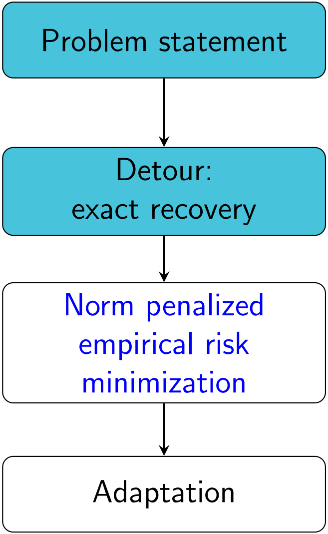
Z given $n \times m$ matrix with $m \gg n$.

Let x^0 be the sparsest solution of $Zx = b^0$.
Basis Pursuit [Chen, Donoho and Saunders (1998)]:

$$x^* := \min \left\{ \|x\|_1 : \; Zx = b^0 \right\}$$

Exact recovery

$$\Gamma(S_0) < \infty \Rightarrow x^* = x^0$$

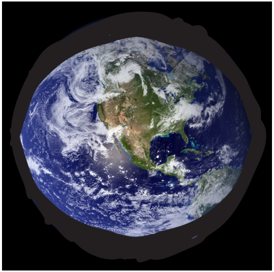


Concepts:

Sparsity
Effective sparsity
Margin curvature
Triangle property

General norms

Let Ω be a norm on $\mathbb{R}^m$



The Ω−world

Norm-regularized empirical risk minimization

$$x^Q := \hat{x} := \arg\min_{x \in \mathcal{X} \subset \mathbb{R}^m} \left\{ \underbrace{f_Q(x)}_{\text{empirical risk}} + \underbrace{\lambda \Omega(x)}_{\text{regularization penalty}} \right\}$$

where

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

Examples of norms

$\boxed{\ell_1\text{-norm:}}$ $\Omega(x) = \|x\|_1 =: \sum_{j=1}^{m} |x_j|$

Examples of norms

ℓ_1-norm: $\Omega(x) = \|x\|_1 =: \sum_{j=1}^{m} |x_j|$

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

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

[Bondell and Reich 2008]

Examples of norms

ℓ_1-norm: $\Omega(x) = \|x\|_1 =: \sum_{j=1}^{m} |x_j|$

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

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

[Bondell and Reich 2008]

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

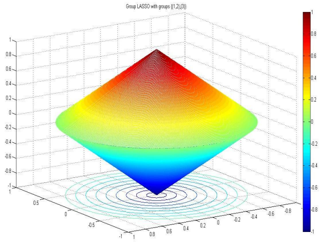
$$\Omega(x) := \sum_{j=1}^{p} \lambda_j |x|_{(j)} \qquad \text{where } |x|_{(1)} \geq \cdots \geq |x|_{(p)}$$

[Bogdan et al. 2013]

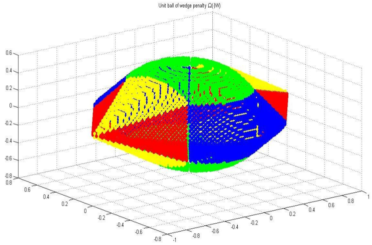
norms generated from cones:

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

[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: $X \in \mathbb{R}^{m_1 \times m_2}$,

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

nuclear norm for matrices: $X \in \mathbb{R}^{m_1 \times m_2}$,

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

nuclear norm for tensors: $X \in \mathbb{R}^{m_1 \times m_2 \times m_3}$,

$\Omega(X) :=$ dual norm of Ω_*
where

$$\Omega_*(W) := \max_{\|u_1\|_2 = \|u_2\|_2 = \|u_3\|_2 = 1} \text{trace}(W^T u_1 \otimes u_2 \otimes u_3), \ W \in \mathbb{R}^{m_1 \times m_2 \times m_3}$$

[Yuan and Zhang 2014]

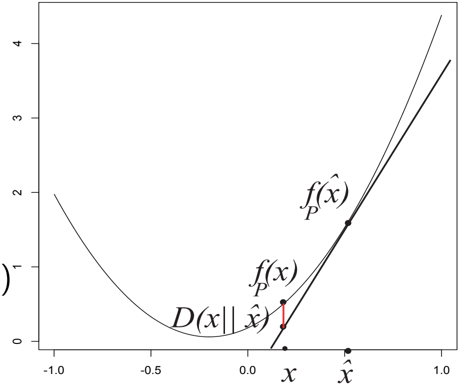
Some concepts

Let $\dot{f}_P(x) := \frac{\partial}{\partial x} f_P(x)$

The Bregman divergence is

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

$$= f_P(x) - f_P(\hat{x}) - \dot{f}_P(\hat{x})^T(x - \hat{x})$$



Definition (Property of f_P) *We have margin curvature G if*

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

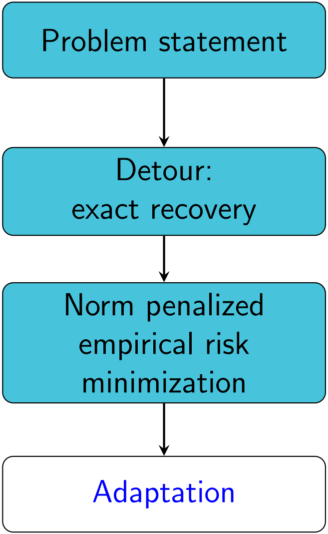
Definition (Property of Ω) *The triangle property holds at x^* if $\exists$ semi-norms Ω^+ and Ω^- such that*

$$\boxed{\Omega(x^*) - \Omega(x) \leq \Omega^+(x - x^*) - \Omega^-(x)}$$

Definition *The effective sparsity at x^* is*

$$\Gamma_*^2(L) := \max\left\{ \left(\frac{\Omega^+(x)}{\tau(x)}\right)^2 : \underbrace{\Omega^-(x) \leq L\Omega^+(x)}_{\text{"cone condition"}} \right\}$$

$L \geq 1$ is a stretching factor.



Concepts:

Sparsity
Effective sparsity
Margin curvature
Triangle property

Norm-regularized empirical risk minimization

$$x^Q := \hat{x} := \arg\min_{x \in \mathcal{X} \subset \mathbb{R}^m} \Big\{ \underbrace{f_Q(x)}_{\text{empirical risk}} + \underbrace{\lambda \Omega(x)}_{\text{regularization penalty}} \Big\}$$

where

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

A sharp oracle inequality

Theorem [vdG, 2016] *Let*

$$\lambda > \lambda_\epsilon \geq \underset{\substack{\uparrow \\ \text{dual norm}}}{\underline{\Omega}_*}\Big(\overset{\substack{\text{this measures how close } Q \text{ is to } P \\ \downarrow}}{(\dot f_Q - \dot f_P)(\hat x)} \Big) \quad \Big(\begin{array}{l} \textit{i.e. remove most} \\ \textit{of the variance} \end{array} \Big)$$

Define

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

Then (recall $\hat x = x^Q$, $x^0 = x^P$ *)*

$$f_P(\hat x) - f_P(x^0) \leq \min_{x^* \in \mathcal{X}} \Big\{ \underbrace{f_P(x^*) - f_P(x^0)}_{\text{“bias”}} + \underbrace{\overset{\substack{H := \text{ convex conjugate of } G \\ \downarrow}}{H(\bar\lambda \Gamma_*(L))}}_{\text{pinch of “variance”}} \Big\}.$$

that is: **Adaptation**

Example: Lasso

$Y \in \mathbb{R}^n$, $Z \in \mathbb{R}^{n \times m}$

Model: $Y = b^0 + \epsilon$

$$f_P(x) := \| b^0 - Zx \|_2^2 / n$$

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

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

Effective sparsity at x^0: $\Gamma_0^2(L) = s_0 / \hat{\phi}^2(L, S_0)$

From the theorem:

with high probability

effective sparsity

$$f_P(\hat{x}) - f_P(x^0) \leq C \times \underbrace{\boxed{\frac{s_0}{\hat{\phi}^2(L, S_0)}}}_{\text{Adaptation}} \frac{1}{n} \times \log m$$

Adaptation

Simulation: Lasso and sorted ℓ_1-norm

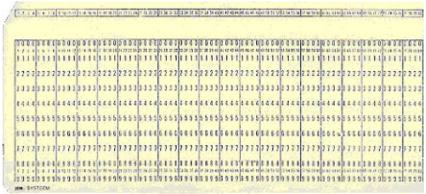
Table

	theoretical λ			cross-validated λ		
	$\|x^0 - \hat{x}\|_1$	$\Omega(x^0 - \hat{x})$	$\|Z(x^0 - \hat{x})\|_{\ell_2}$	$\|x^0 - \hat{x}\|_1$	$\Omega(x^0 - \hat{x})$	$\|Z(x^0 - \hat{x})\|_{\ell_2}$
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

Simulation: Lasso and sorted ℓ_1-norm

Table

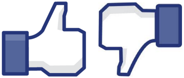
	theoretical λ			cross-validated λ		
	$\|x^0 - \hat{x}\|_1$	$\Omega(x^0 - \hat{x})$	$\|Z(x^0 - \hat{x})\|_{\ell_2}$	$\|x^0 - \hat{x}\|_1$	$\Omega(x^0 - \hat{x})$	$\|Z(x^0 - \hat{x})\|_{\ell_2}$
srSLOPE	4.50	0.49	7.74	7.87	1.09	7.68
srLASSO	8.48	0.89	29.47	7.81	0.85	9.19



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

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

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

Let $Y_i \in \Big\{$  $\Big\}$

Model:

log-odds$(Y_i) = x_i^0 = \mathrm{trace}(Z_i X^0)$

$$f_Q(X) := -\frac{1}{n}\sum_{i=1}^{n} Y_i \mathrm{trace}(Z_i X) + \sum_{j,k} d(X_{j,k})/(m_1 m_2),$$

given

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

Dual norm: operator norm

Margin semi-norm: $\tau^2(X) = \|X\|_2^2/(m_1 m_2)$

Margin curvature:
$G(u) = u^2/(2c m_1 m_2)$
$\Rightarrow H(v) = c m_1 m_2 v^2/2$

Effective sparsity: $\Gamma_0^2(L) = 3s_0$

From the theorem:

for $m_1 \geq m_2$

and $\lambda = C_0 \frac{1}{\sqrt{nm_2}}(\sqrt{\log m_1 + \log(1/\alpha)/m_1}$,

with probability at least $1-\alpha$

$$f_P(\hat{X}) - f_P(X^0) \leq C \times \left(\frac{s_0 m_1 \log(m_1)}{n} \right).$$

Adaptation

Example: Sparse PCA

- $Y_1, \ldots, Y_n$ sample from distribution P on $\mathbb{R}^m$ with covariance matrix Σ_P
- $\Sigma_Q := Y^T Y / n$
- $f_P(x) := \|\Sigma_P - xx^T\|_2^2$, $f_Q(x) := \|\Sigma_Q - xx^T\|_2^2$
- $\Omega := \| \cdot \|_1$

From the theorem:

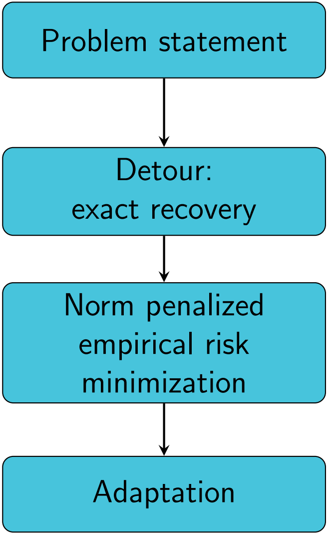
Assume ...

Then with $\lambda = C_0 \sqrt{\log m / n}$, w.h.p.[1]

$$f_P(\hat{x}) - f_P(x^0) \leq C_1 \frac{s_0 \log m}{n}$$

Adaptation

[1] this means: with high probability



Concepts:

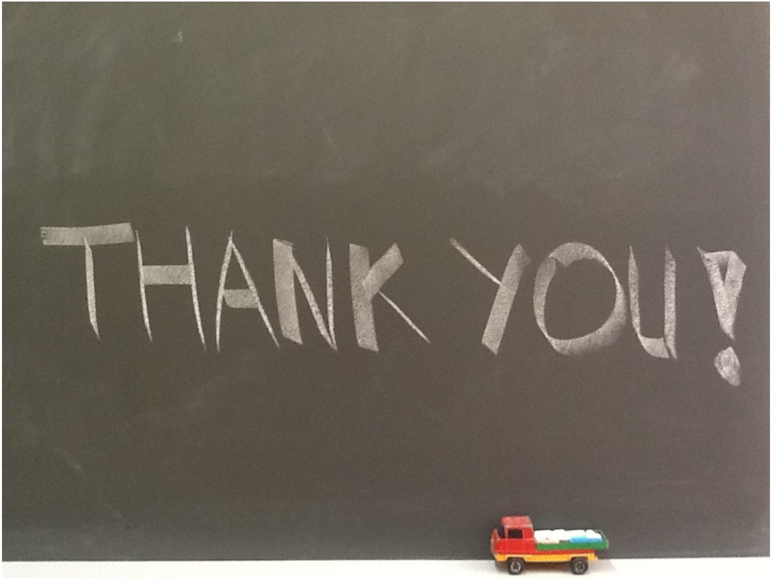
Sparsity
Effective sparsity
Margin curvature
Triangle property

Conclusion

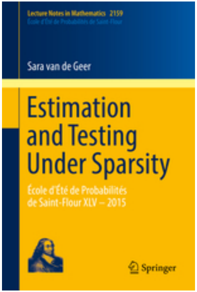
norms with the triangle property

lead to **Adaptation**

for general loss and assuming margin curvature


THANK YOU!

See:



and its references