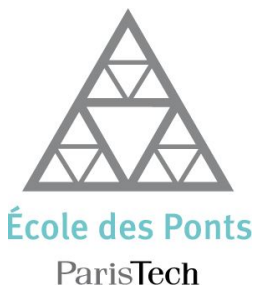
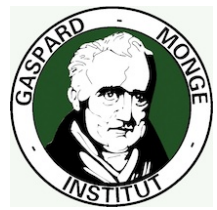
Sparse methods and compressed sensing



Guillaume Obozinski

Ecole des Ponts - ParisTech



CGL workshop

Athens, Sep 29th-Oct 2nd 2013

Why sparsity? Signal processing point of view

Signal processing point of view

- High-dimensional signals
- But with simple regular/sparse latent structure
- Natural images do not fill the space of possible images. Locally like a low dimensional manifold.
- Well approximated on a wavelet basis with a small number of coefficients.
- Be able to process, compress, denoise, restore those signals by leveraging their low-dimensional structure.

Why sparsity? Machine learning point of view

Machine learning point of view

- Learn to perform a task (classification, regression, ranking, etc) from a vector representation of an object (features)
- Too many features (high dimension)
- Leads to overfitting
- Maybe only a few features matter

Classical supervised learning setup (ERM)

- Data: $(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$
- $f \in \mathcal{H}$ function to learn
- Loss function: $\ell : (y, a) \mapsto \ell(y, a)$
  - e.g. $\ell(y, a) = \frac{1}{2}(y - a)^2$, logistic loss, hinge loss, etc.

Empirical Risk Minimization

$$\min_{f \in \mathcal{H}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)}_{\text{Empirical Risk}} + \underbrace{\lambda \|f\|_{\mathcal{H}}^2}_{\text{Regularization}}$$

- $\mathcal{H}$ typically an RKHS
- λ: regularization coefficient
- λ controls the complexity of the function that we are willing to learn for a given amount of data.

Learning linear functions

Restricting to linear functions $f_w : x \mapsto w^\top x_i$

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(w^\top x_i, y_i) + \frac{\lambda}{2} \|w\|^2$$

- For the square loss $\rightarrow$ ridge regression
- **Issue:** number of features p typically large compared to the amount of data

Alternative to regularization provided by sparsity

Reducing the number of features entering the models yields

- another way to control model complexity
- more interpretable models (very important in biomedical applications)
- computationally efficient algorithms

Outline of the lecture

1. Algorithms
   - Greedy
   - Convex formulations
2. Compressed sensing
3. Restricted Isometry Property, Mutual Incoherence
4. Group sparsity and beyond.
5. Atomic norm and a geometric point of view

The simplest sparse model

- $w \in \mathbb{R}^p$ sparse signal: $\|w\|_0 \leq k$

Observe:

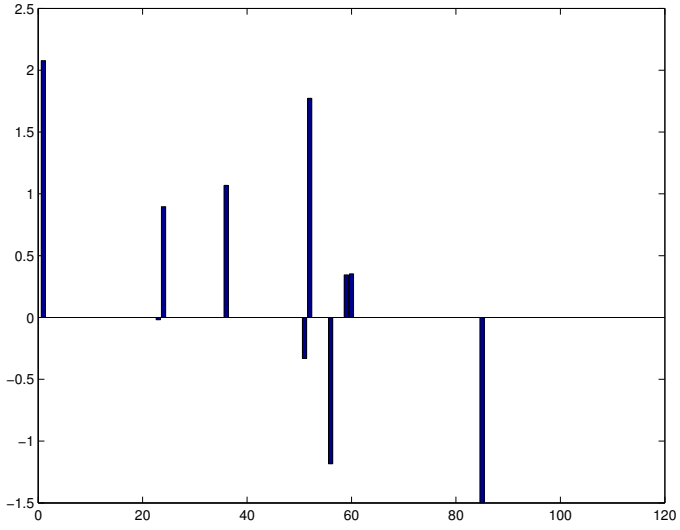
$$y = w + \epsilon,$$

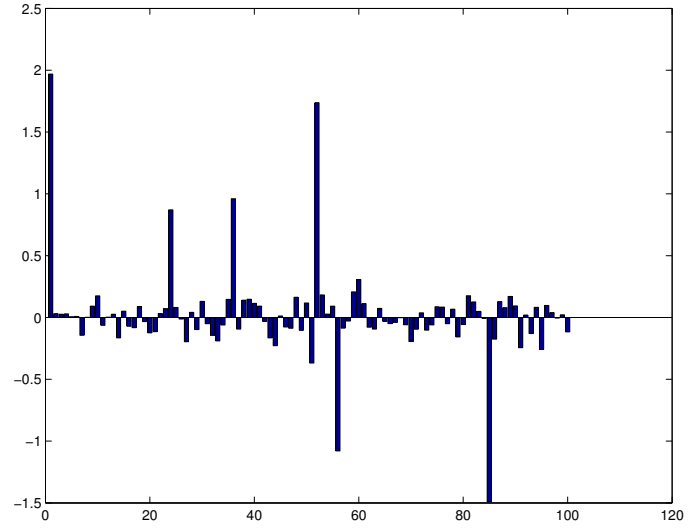
with

$$\epsilon_i \quad \text{i.i.d.} \quad \mathbb{E}[\epsilon_i] = 0,\ \mathrm{Var}(\epsilon_i) = \sigma^2.$$

Denoising a sparse vector

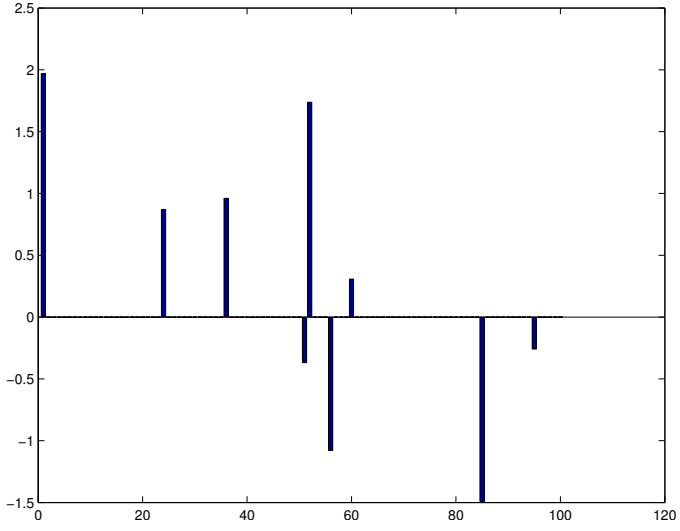
Raw signal

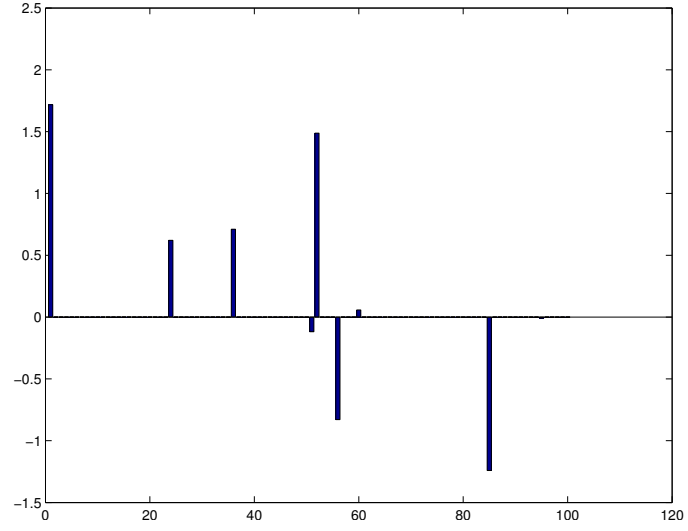




Noisy signal

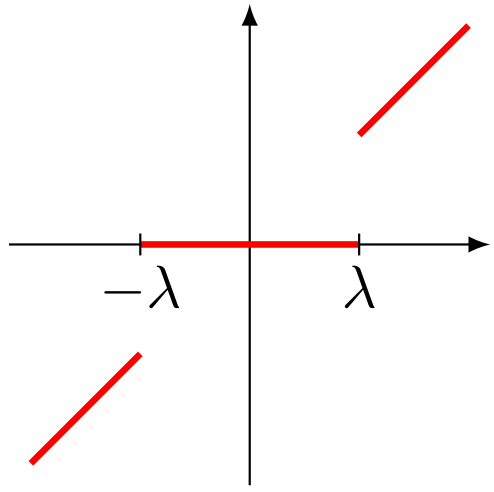
Hard thresholding

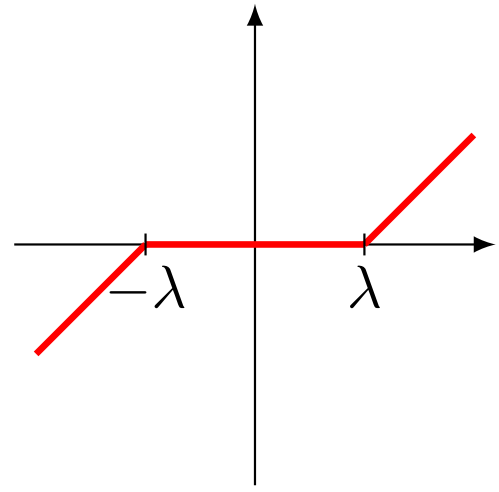




Soft thresholding

Hard-thresholding *vs* Soft-thresholding

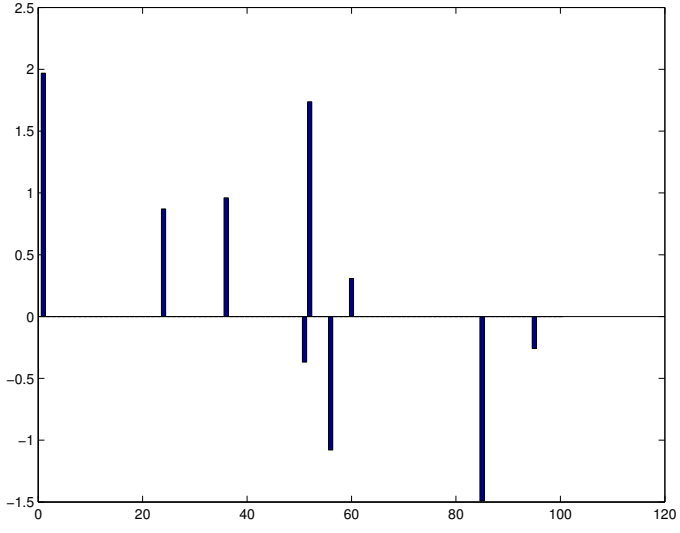


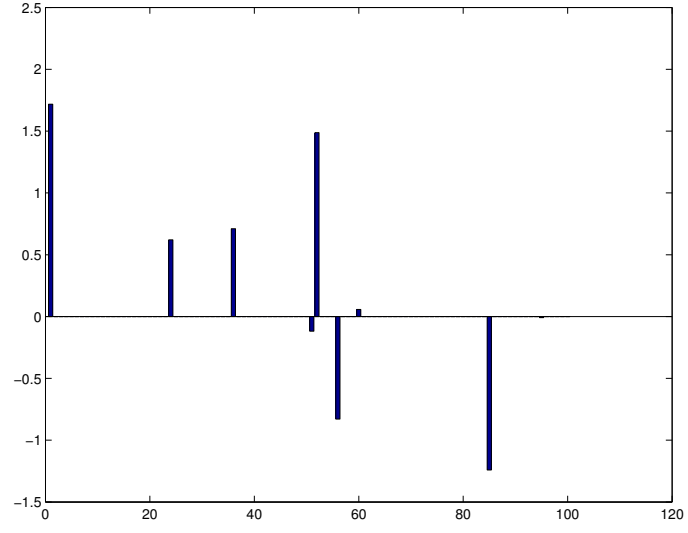


$t \mapsto 1_{\{|t|>\lambda\}}\, t$

$t \mapsto (|t| - \lambda)_+ \,\mathrm{sign}(t)$

Hard thresholding





Soft thresholding

A sparse signal

- $y \in \mathbb{R}^n$ is the signal
- $X \in \mathbb{R}^{n \times p}$ is some overcomplete basis
- w is the sparse representation of the signal

Find w sparse such that

$$y = Xw$$

Classical signal processing formulation of the problem

$$\min \|w\|_0 \quad \text{s.t.} \quad y = Xw.$$

Problem: there is noise... and noise is not sparse

$$\min \|w\|_0 \quad \text{s.t.} \quad \|y - Xw\|_2 \leq \epsilon.$$

$$\min \|y - Xw\|_2^2 \quad \text{s.t.} \quad \|w\|_0 \leq k$$

$$\min \frac{1}{2}\|y - Xw\|_2^2 + \lambda \|w\|_0$$

These problems are NP-hard.

Approaches

Greedy Methods

- Matching Pursuit (MP)
- Orthogonal Matching Pursuit (OMP)
- Least-square OMP
- CoSamp

Relaxation Methods

- Lasso/Basis Pursuit
- Dantzig Selector

Bayesian Methods

- Spike and Slab priors (ARD)
- Empirical Bayes

Greedy Methods

Greedy methods

Principle: $\hat{\beta}$ is estimated by increasing the support greedily. At each iteration

1. **Selection step:** A new coordinate is included in the support of $\hat{\beta}$
2. **Fitting step:** The new coefficient and possibly old ones are re-optimized

Orthogonal Matching Pursuit

$$\min \|w\|_0 \quad \text{s.t.} \quad y = Xw.$$

$$\min \frac{1}{2}\|y - Xw\|_2^2 + \lambda \|w\|_0$$

Initialization:

- $\hat{S} = \varnothing$ (estimate of support)
- $r \leftarrow y$ (residuals)

Repeat:

1. **Selection Step:**
   - $i \leftarrow \arg\max_{i'} |\langle x_{i'}, r \rangle|,$
   - $\hat{S} \leftarrow \hat{S} \cup \{i\}$
2. **Fitting Step:**
   - $\hat{w}_{\hat{S}} \leftarrow \arg\min_{w_{\hat{S}}} \|y - X_{\hat{S}}\, w_{\hat{S}}\|_2^2, \qquad r \leftarrow y - X_{\hat{S}}\, \hat{w}_{\hat{S}}$

Matching Pursuit

$$\min \|w\|_0 \quad \text{s.t.} \quad y = Xw.$$

$$\min \frac{1}{2}\|y - Xw\|_2^2 + \lambda \|w\|_0$$

Initialization:

- $\hat{S} = \varnothing$ (estimate of support)
- $r \leftarrow y$ (residuals)

Repeat:

1. **Selection Step:**
   - $i \leftarrow \arg\max_{i'} |\langle x_{i'}, r \rangle|,$
   - $\hat{S} \leftarrow \hat{S} \cup \{i\}$
2. **Fitting Step:**
   - $\hat{w}_i \leftarrow \arg\min_{w_i} \|r - x_i\, w_i\|_2^2, \qquad r \leftarrow r - x_i\, \hat{w}_i$

Least square matching pursuit

$$\min \|w\|_0 \quad \text{s.t.} \quad y = Xw.$$

$$\min \frac{1}{2}\|y - Xw\|_2^2 + \lambda \|w\|_0$$

Initialization:

- $\hat{S} = \varnothing$ (estimate of support)
- $r \leftarrow y$ (residuals)

Repeat:

1. **Selection Step:**
   - $i \leftarrow \arg\min_{i'} \min_{w_{\hat{S}\cup\{i'\}}} \|y - X_{\hat{S}\cup\{i'\}}\, w_{\hat{S}\cup\{i'\}}\|_2^2,$
   - $\hat{S} \leftarrow \hat{S} \cup \{i\}$
2. **Fitting Step:**
   - $\hat{w}_{\hat{S}} \leftarrow \arg\min_{w_{\hat{S}}} \|y - X_{\hat{S}}\, w_{\hat{S}}\|_2^2, \qquad r \leftarrow y - X_{\hat{S}}\, \hat{w}_{\hat{S}}$

Convex formulations

A convex relaxation...

- Empirical risk: for $w \in \mathbb{R}^p$,

$$L(w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

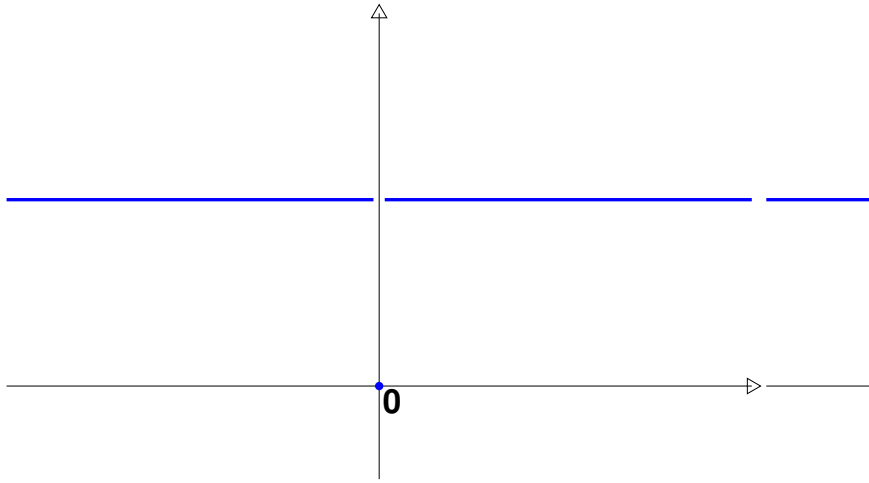
$$|\mathrm{Supp}(w)| = \sum_{i=1}^{n} 1_{\{w_i \neq 0\}}$$

- Support of the model:

$$\mathrm{Supp}(w) = \{i \mid w_i \neq 0\}.$$

Penalization for variable selection

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \, |\mathrm{Supp}(w)|$$



Lasso (Tibshirani, 1996)

$$\min_{w \in \mathbb{R}^d} L(w) + \lambda \|w\|_1$$

Formulations through convex programs

Basis Pursuit (Chen et al., 1998)

$$\min_w \|w\|_1 \quad \text{s.t.} \quad y = Xw$$

Basis Pursuit ("noisy" setting)

$$\min_w \|w\|_1 \quad \text{s.t.} \quad \|y - Xw\|_2 \leq \eta$$

Lasso (Tibshirani, 1996)

$$\min_w \frac{1}{2n}\|y - Xw\|^2 + \lambda\|w\|_1$$

Dantzig Selector (Candès and Tao, 2007)

$$\min_w \|w\|_1 \qquad \text{s.t.} \qquad \|X^\top(y - Xw)\|_\infty \leq \lambda$$

Remarks

- Minima not necessarily unique
- Dantzig Selector $\rightarrow$ linear program
- Theoretical guarantees for the Lasso and $\|X^\top(y - Xw)\|_\infty \leq \lambda$

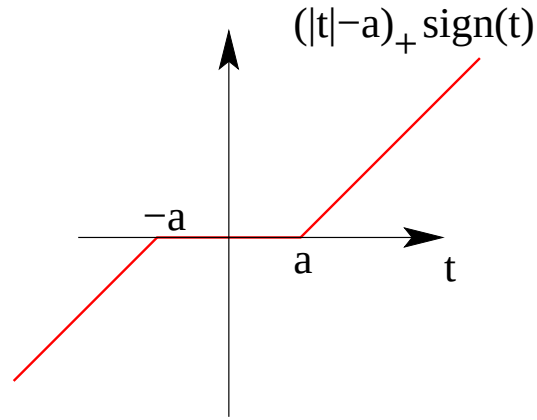
Lasso with orthogonal design

- Assume $\frac{1}{n} X^\top X = I$.
- Then solving the Lasso is equivalent to solving

$$\min_{w} \frac{1}{2n} \|X^\top y - w\|_2^2 + \lambda \|w\|_1$$

$$\min_{v \in \mathbb{R}} \frac{1}{2} v^2 - vt + a|v| \quad \rightarrow \quad v^* = (|t| - a)_+ \, \mathrm{sign}(t)$$

cf $\min_w \frac{1}{2n} \left(w^\top X^\top X w - 2 w^\top X^\top y + \|y\|^2 \right) + \lambda \|w\|_1$

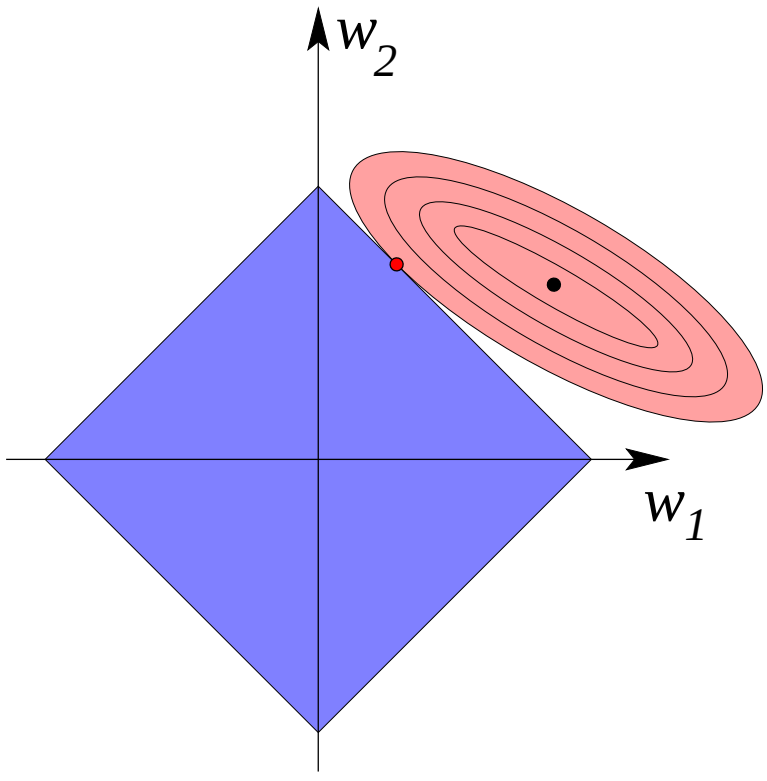


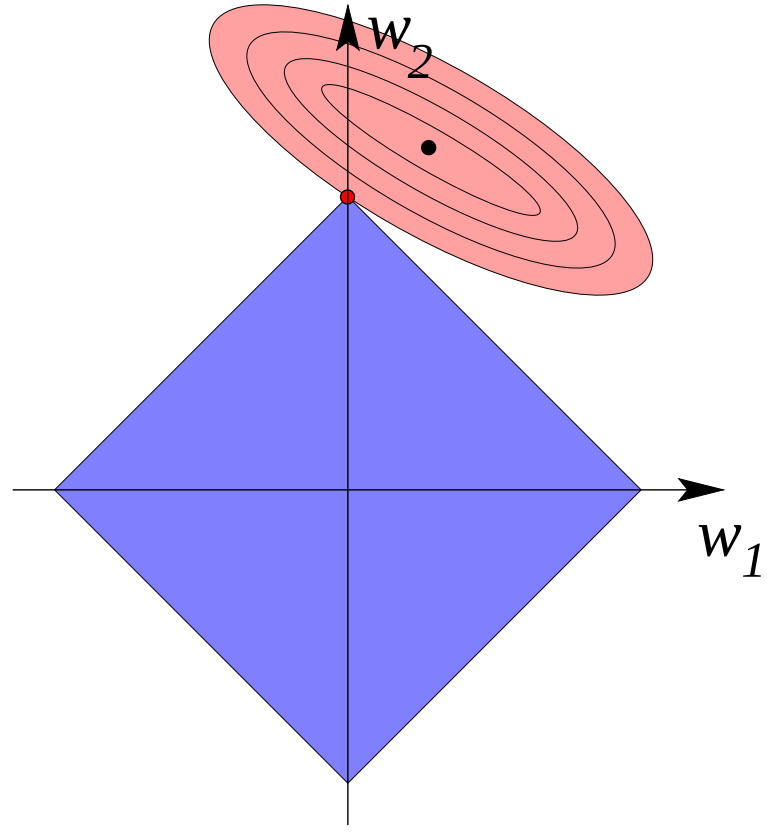
$\rightarrow$ Soft-thresholding:

$$\begin{aligned} \hat{w}_j &= \mathrm{ST}_\lambda\left(\frac{1}{n} X_j^\top y \right) \\ &= \mathrm{ST}_\lambda\left(w_j^\star + \frac{1}{n} X_j^\top \varepsilon \right) \end{aligned}$$

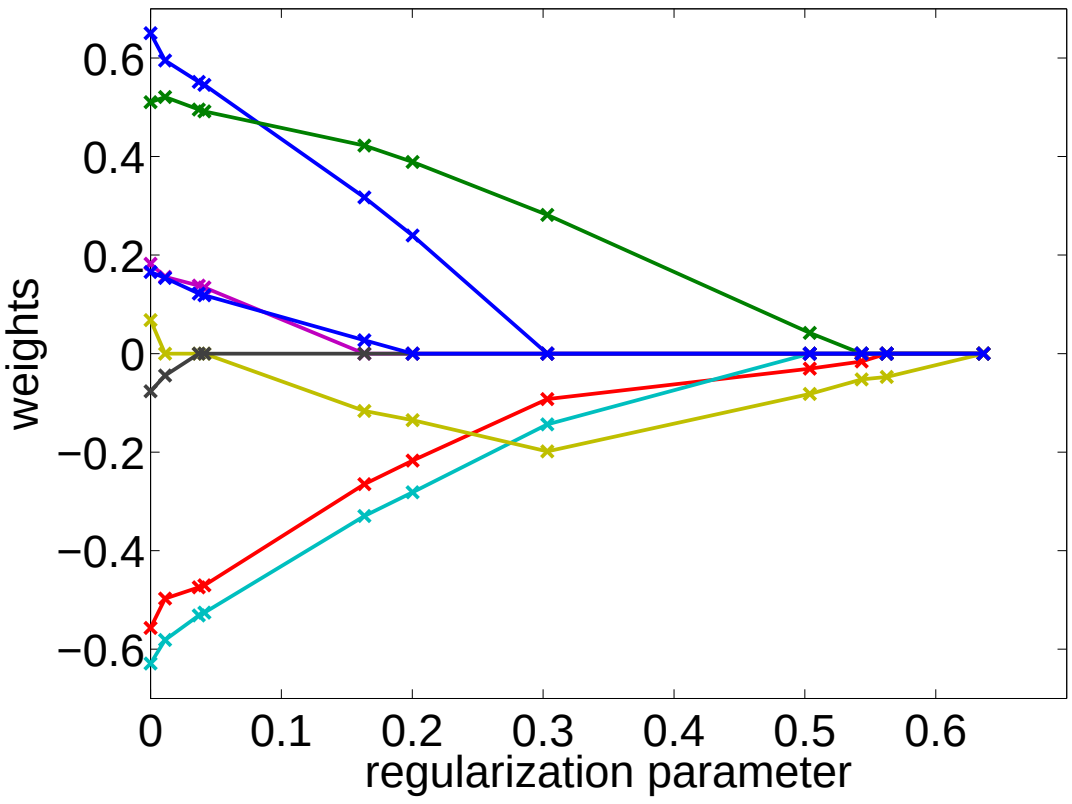
Why ℓ_1-norm constraints leads to sparsity?

- Example: minimize quadratic function $Q(w)$ subject to $\|w\|_1 \leqslant T$.
  - coupled soft thresholding
- Geometric interpretation
  - NB : penalizing is "equivalent" to constraining





Piecewise linear paths



Optimization algorithms

Generic approaches

- For the Lasso, with interior point methods.
- Subgradient descent

Efficient first order methods

- Coordinate descent methods
- Proximal methods
- Reweighted ℓ_2 methods (esp. for structured sparse methods)

Active set methods

- For the Lasso: LARS algorithm
- In general: meta-algorithms to combine with methods above

Theoretical guarantees

Types of theoretical guarantees

What guarantees could we give about sparse methods?

Wish list

We would like that:

- If there is no noise, the sparse signal is recovered exactly
- If there is noise:
  1. The exact support of w is recovered
  2. $\|\hat{w} - w\|_2$ is small
  3. In ML context, $\mathbb{E}\left[\ell(\hat{w}^\top X, Y)\right]$ is small.

Can give guarantees

- if the columns of X are sufficiently decorrelated.
- One then says that X is **incoherent**

Spark of a matrix Donoho and Elad (2003)

- spoon + fork = ? $\rightarrow$ spork
- sparse + rank = ? $\rightarrow$ spark

Definition (Spark $\tilde{\sigma}$ of X)

Smallest number of columns of X that are not independent.

$\Leftrightarrow$ Smallest k such that equivalently

- there exists a set of k columns that are not of full rank
- $\lambda_{\min}(k) = 0$
- $\delta_k \geq 1$

Consequence:

$$\boxed{\tilde{\sigma} \geq 1 + \frac{1}{\mu}}$$

- using $\delta_k \leq \mu\,(k-1)$

Recovery

$$\boxed{\hat{\beta} = \arg\min \|\beta\|_0 \quad \text{s.t.} \quad y = X\beta}$$

- When does it work?

Lemma (Identifiability)

If $\|\hat{\beta}\|_0 < \frac{\tilde{\sigma}}{2}$ *then it is the unique minimal solution.*

Proof.

If there is another solution β such that $\|\beta\|_0 < \frac{\tilde{\sigma}}{2}$,
then $\|\hat{\beta} - \beta\|_0 \leq \|\hat{\beta}\|_0 + \|\beta\|_0 < \tilde{\sigma} \quad \Rightarrow \quad \beta = \hat{\beta}$ □

Compressed Sensing

The point of view of compressed sensing

- Reference is Shannon-Nyquist theory
- Simplified view
  - Signal: $w \in \mathbb{R}^p$
  - Sensing matrix $X \in \mathbb{R}^{n \times p}$
  - Measurement vector $y = Xw \in \mathbb{R}^n$

Idea w has some latent structure that allows to compress it and that can be exploited to design a sampling scheme that beats the Nyquist rate.

Adaptive vs non-adaptive measurements

Assume that w is "compressible", e.g., either

- sparse with support S such that $|S| = k$
- has quickly decaying coefficients (e.g., ℓ_p for $p \leq 1$)

Then w can be encoded/compressed *adaptively* by resp. storing

- the indices of the non-zero coefficients and their values (lossless)
- the indices and values of the k largest coefficient (lossy)

Is it possible to make a small number of non-adaptive measurements that achieves the same compression?

Idea The compressed sensing device will not know the signal structure/sparsity to begin with...

- $Xw \in \mathbb{R}^n$ where X is fixed independently of w
- Is it possible to encode w with only n measurement and $n \ll p$?

A more realistic view of compressed sensing

- Signal: $Dw + z$ with $D \in \mathbb{R}^{m \times p}$ and $w \in \mathbb{R}^p$ is sparse (compressible) and z some noise
- Sensing matrix $X \in \mathbb{R}^{n \times p}$
- $y = X(Dw + z) + \epsilon$

Goal: Design X to recover

- Dw or
- w

as well as possible

Idea: Each of the rows of X is a non-adaptive measurement acquired as a single value

Restricted Isometry Property (Candes and Tao, 2005)

Is it possible to encode w with only n measurements?

Answer If $n = p$ take X to be

- an invertible transformation
- even better an *isometry*

But we want $n \ll p$!

Surprise! It is still possible for X to have a restricted *isometry* property.

Definition (Restricted Isometry Property)

$X \in \mathbb{R}^{p \times n}$ satisfies a restricted isometry property for k-sparse vectors of $\mathbb{R}^p$ if there exists $\delta_k \in [0, 1)$ such that,
for all $w \in \mathbb{R}^p$ such that $|\text{supp}\,(w)\,| \leq k$

$$(1 - \delta_k)\,\|w\|_2^2 \;\leq\; \|Xw\|_2^2 \;\leq\; (1 + \delta_k)\,\|w\|_2^2$$

δ_k is called the RIP constant.

A second RIP constant and Mutual Incoherence

δ_k and $\theta_{k,k'}$ are RIP constants for matrix X, if for all β and β' with support respectively S and S' such that $|S| = k$, $|S'| = k'$ and $S \cap S' = \varnothing$, we have

$$(1 - \delta_k) \|\beta\|_2^2 \leq \|X\beta\|_2^2 \leq (1 + \delta_k) \|\beta\|_2^2$$

and

$$\langle X\beta, X\beta' \rangle \leq \theta_{k,k'} \|\beta\|_2 \|\beta'\|_2$$

Mutual Incoherence: $\boxed{\mu = \delta_{1,1} = \max_{i \neq j} |\langle x_i, x_j \rangle|}$

Properties:

$$\boxed{\theta_{k,k'} \leq \delta_{k+k'} \leq \theta_{k,k'} + \max(\delta_k, \delta_{k'})}$$

$\boxed{\theta_{k,k'} \leq \sqrt{k\,k'}\,\mu}$ and $\boxed{\delta_k \leq \mu\,(k-1)}$

Properties of high-dimensional spaces

- Very surprising and different than in low dimensional spaces.
- In dimension n, the number of directions that are *almost orthogonal*

  with "almost orthogonal" meaning $\quad |\langle x, y \rangle| \leq \epsilon$

  ... is *almost exponential* in n !!!
- For any m points in $\mathbb{R}^p$ it is possible to map them to $\mathbb{R}^n$ with an arbitrarily small distortion of the distances between the points for $n = O(\log(p))$.

Johnson-Lindenstrauss Lemma

Lemma (Johnson-Lindenstrauss)

- *For any* $\epsilon > 0$, *for any integer* m,
- *let* n *be such that*

$$n > \frac{8 \log m}{\epsilon^2} \left(1 - \frac{2}{3}\epsilon\right)^{-1}$$

Then for any set $\mathcal{M}$ *of* m *points, there exists a Lipschitz map* $f : \mathbb{R}^p \to \mathbb{R}^n$ *such that*

$$\forall u, v \in \mathcal{M}, \qquad (1-\epsilon)\|u-v\|_2^2 \leq \|f(u)-f(v)\|_2^2 \leq (1+\epsilon)\|u-v\|_2^2,$$

and this map can be found in randomized polynomial time.

Comments

- f is a *quasi-isometry*
- f can be constructed as a *linear map* !
- f can be obtained by random sampling !!

RIP for random matrices

Theorem (Baraniuk et al. (2008))

Let $X \in \mathbb{R}^{n \times p}$ be a random Gaussian matrix with $X_{ij} \sim \mathcal{N}(0, \frac{1}{n})$. Then, there exist constants $c_1, c_2 > 0$ depending only on δ such that, with probability $\geq 1 - 2e^{-c_2 n}$, the RIP holds for X

- *for the prescribed δ*
- *for any k satisfying $n \geq c_1 k \log(\frac{p}{k})$*

- We cannot guarantee/check that the matrix has the RIP property
- We need to take n much larger ($n > k^2 \log p$) to have the MI property

Dantzig Selector and Basis Pursuit (noisy) under RIP

Theorem (Cai et al. (2010))

- *Assume* $\delta_{1.25\,k} + \theta_{k,\,1.25k} < 1$

1. *For Basis Pursuit with the constraint* $\|Y - X\hat{\beta}\|_2 \leq \eta$
   - *Assume that* $\|Y - X\beta\|_2 \leq \epsilon$

   Then

$$\boxed{\|\hat{\beta} - \beta\|_2 \leq (\eta + \epsilon)\,\frac{\sqrt{2 + 2\delta_{1.25k}}}{1 - \delta_{1.25\,k} + \theta_{k,\,1.25k}}}$$

2. *For the Dantzig selector with the constraint* $\|X^\top(Y - X\hat{\beta})\|_\infty \leq \eta$
   - *Assume that* $\|X^\top(Y - X\beta)\|_\infty \leq \epsilon$

   Then

$$\boxed{\|\hat{\beta} - \beta\|_2 \leq (\eta + \epsilon)\,\frac{\sqrt{10\,k}}{1 - \delta_{1.25\,k} - \theta_{k,\,1.25k}}}$$

Conditions on the design

- **Restricted Isometry Property (RIP)**

$$\sqrt{1-\delta_k}\|w\| \leq \|Xw\| \leq \sqrt{1+\delta_k}$$

Subsets of size k of the columns of X should be close to an orthonormal system.

- **Mutual Incoherence Property (MIP)**

$$\max_{i \neq j} |x_i^\top x_j| < \mu$$

Very close to forming an orthogonal basis.

- **Irrepresentable condition (IC)**

$$\|\Sigma_{S^c S} \Sigma_{SS}^{-1}\|_\infty \leqslant 1 - \gamma$$

with $\Sigma_{SS'} = X_S^\top X_S$.

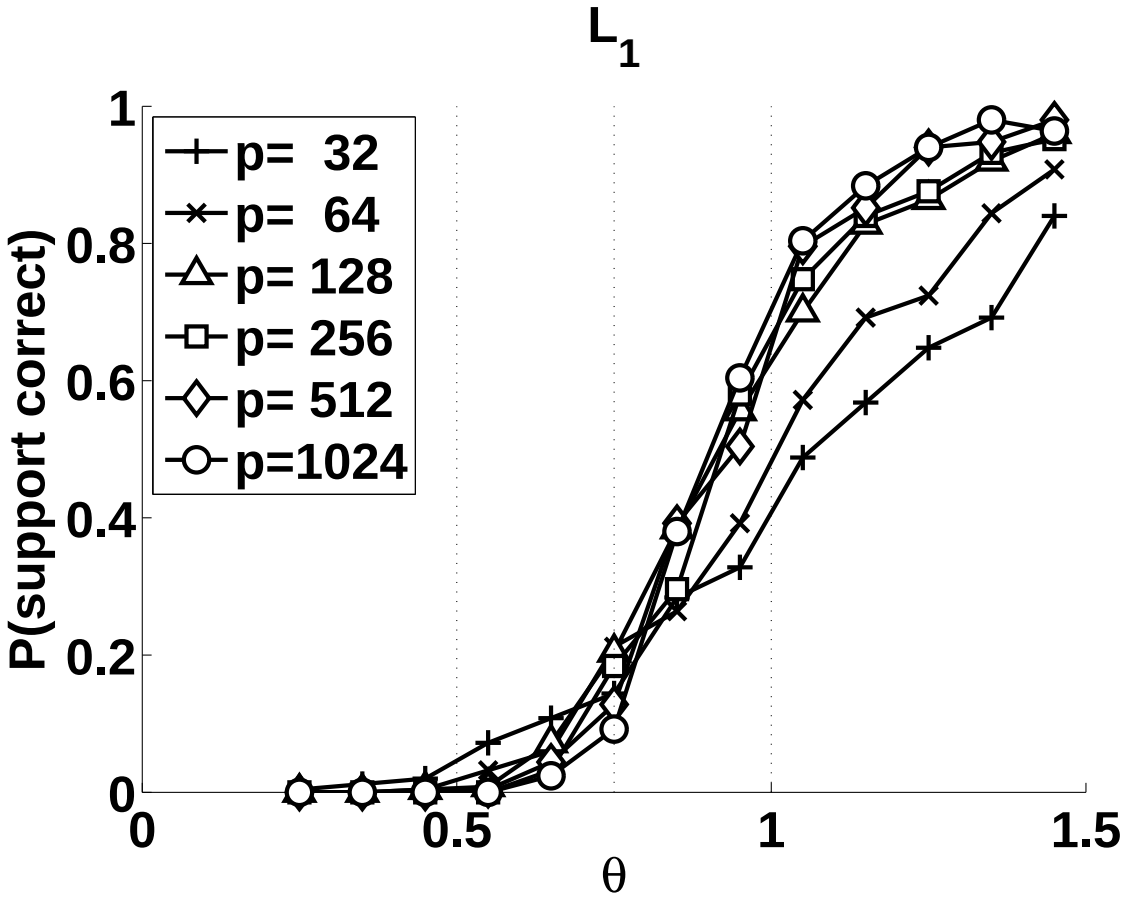
- **Restricted Eigenvalue condition (RE)**

$$\kappa(k)^2 = \min_{|S| \leqslant k} \quad \min_{\Delta,\ \|\Delta_{S^c}\|_1 \leqslant \|\Delta_S\|_1} \frac{\Delta^\top X^\top X \Delta}{\|\Delta_S\|_2^2} > 0$$

Phase transition for support recovery (Wainwright, 2009)

For a random Gaussian design $X \in \mathbb{R}^{n,p}$ with X_{ij}, i.i.d. $\mathcal{N}(0, \frac{1}{n})$, for λ well chosen and $\min_{i \in S} |w_i|$ not too small:

- If $n > 2k \log(p) \rightarrow$ w.h.p support recovered
- If $n < 2k \log(p) \rightarrow$ w.h.p. support not recovered



$$\theta = \frac{n}{2k \log(p)}$$

Comparing Lasso and other strategies for linear regression

- Compare:

  Ridge regression: $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|y - Xw\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$

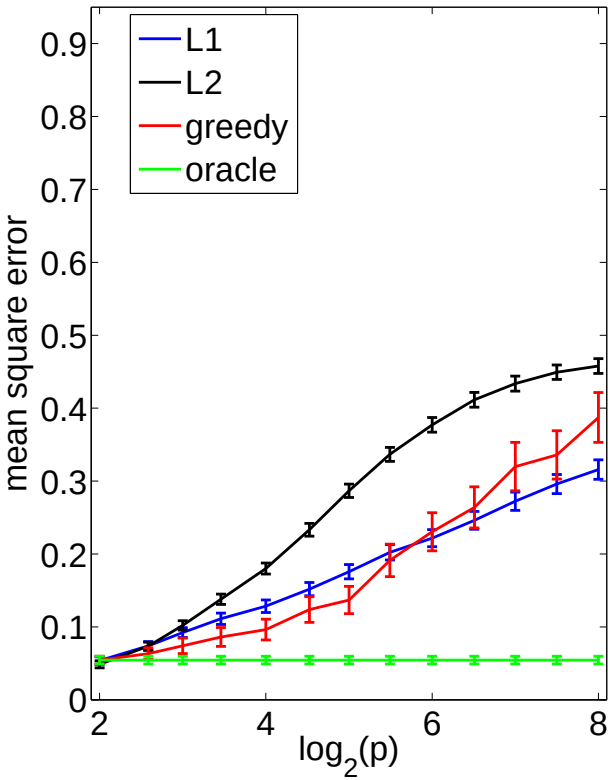
  Lasso: $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|y - Xw\|_2^2 + \lambda\|w\|_1$

  OMP/FS: $\min_{w \in \mathbb{R}^p} \frac{1}{2}\|y - Xw\|_2^2 + \lambda\|w\|_0$

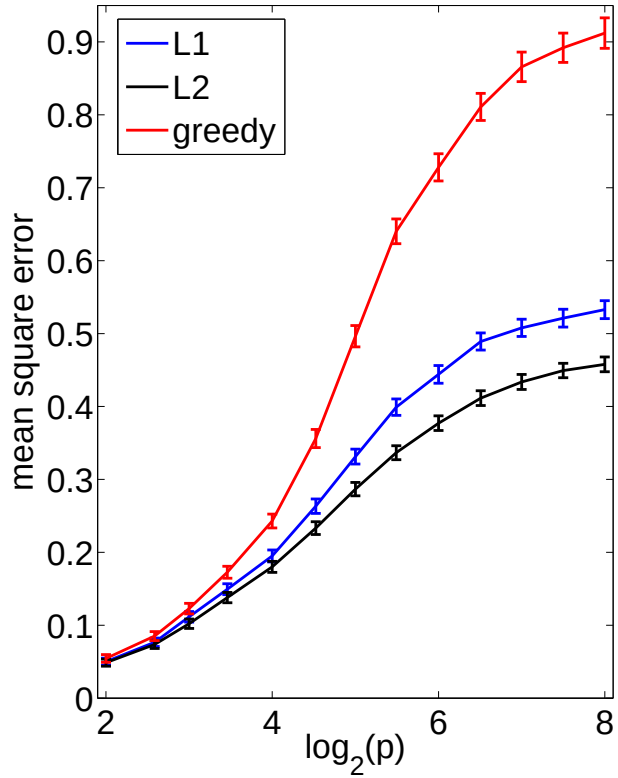
- Each method builds a path of solutions from 0 to ordinary least-squares solution
- Regularization parameters selected on the test set

Simulation results

- i.i.d. Gaussian design matrix, $k = 4$, $n = 64$, $p \in [2, 256]$, SNR $= 1$
- Note stability to non-sparsity and variability



Sparse



Rotated (non sparse)

Advantages and Drawbacks of ℓ_1 *vs* ℓ_0 penalization

Advantages

- The solution $\alpha(x)$ is a continuous (differentiable on the support) function of the data x.
- The ℓ_1-norm is more robust to violation of the sparsity assumption.
- It controls the influence of spuriously introduced variables (like $\ell_0+\ell_2$)
- The convex formulation leads to principled algorithms that generalize well to new situations and natural theoretical analyses.

Drawbacks

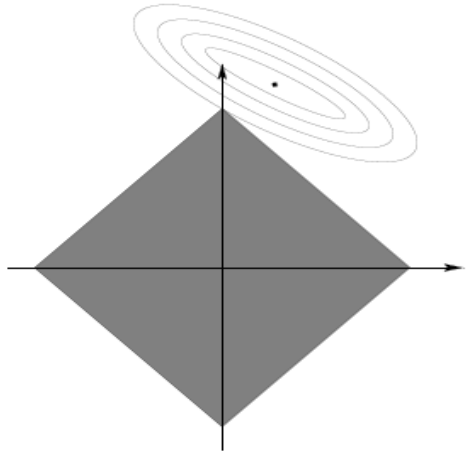
- It introduces an estimation bias which leads to the selection of two many variables if ignored.
- Some of the ℓ_0 algorithms are simpler.

Group sparsity, matrix sparsity

From ℓ_1-regularization...

- $$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - w^\top x^{(i)})^2 + \lambda \|w\|_1$$

  with $\|w\|_1 = \sum_{j=1}^{p} |w_j|$.



...to penalization with grouped variables

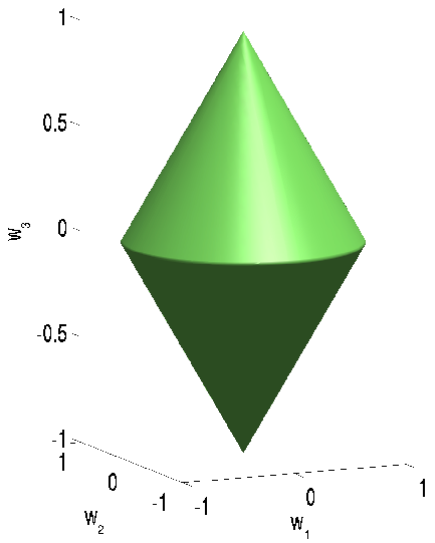
Assume that $\{1, \ldots, p\}$ is **partitioned** into m groups $G_1, \ldots, G_m$

$$w = (w_{G_1}, \ldots, w_{G_m})^\top \quad \text{and} \quad x = (x_{G_1}, \ldots, x_{G_m})^\top$$

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(w^\top x^{(i)}, y^{(i)}) + \lambda \sum_{j=1}^{m} \|w_{G_j}\|$$

Group Lasso (Yuan and Lin, 2006)

- $\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - w^\top x^{(i)})^2 + \lambda \sum_{j=1}^{m} \|w_{G_j}\|$
- The ℓ_1/ℓ_2 norm: $\Omega(w) := \sum_{G \in \mathcal{G}} \|w_G\|_2$
- Unit ball in $\mathbb{R}^3$: $\|(w_1, w_2)\| + \|w_3\| \leq 1$
- Some entire groups set to 0
- No zero within groups



ℓ_1/ℓ_q-regularization

- Can also consider ℓ_1/ℓ_∞-norm
  - More non-differentiabilities

Applications

- Group of nominal variables (dummy binary variables)
- Learn sums of polynomial functions:
  $f(x) = f(x_1) + \ldots + f(x_p)$

$$\min_w \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j,k} w_{jk} x_j^{(i)^k} - y^{(i)} \right)^2 + \sum_{j=1}^{p} \|(w_{j1}, \ldots, w_{jK})\|_2$$

  - j: variables
  - i: observations
  - k: degree of monomial

Trace norm aka Nuclear norm

- Let $M \in \mathbb{R}^{p \times n}$ be a rectangular matrix
- $M = USV^\top$ its singular value decomposition, with U and V orthonormal bases and $S = \mathrm{Diag}(\sigma)$ a diagonal matrix.

The function

$$M \mapsto \sum_k \sigma_k$$

is a unitarily invariant norm.
It is easy to see that its unit ball is exactly

$$\mathrm{convHull}\big(\{uv^\top \mid u \in \mathbb{S}^{p-1}, v \in \mathbb{S}^{n-1}\}\big)$$

Geometric point of view with atomic norms

(Chandrasekaran et al., 2012)

Atomic Norm (Chandrasekaran et al., 2012)

Given a set of atoms $\mathcal{A}$.

Consider the gauge of the convex hull of $\mathcal{A}$

$$\|x\|_{\mathcal{A}} = \inf\{t > 0 \mid x \in t\,\mathrm{conv}(\mathcal{A})\}.$$

(NB: This is really a norm if $\mathcal{A}$ is centrally symmetric and spans $\mathbb{R}^p$)

$$\|x\|_{\mathcal{A}} = \inf\left\{ \sum_{a\in\mathcal{A}} c_a \mid x = \sum_{a\in\mathcal{A}} c_a a, \quad c_a > 0,\ \forall a \in \mathcal{A} \right\}$$

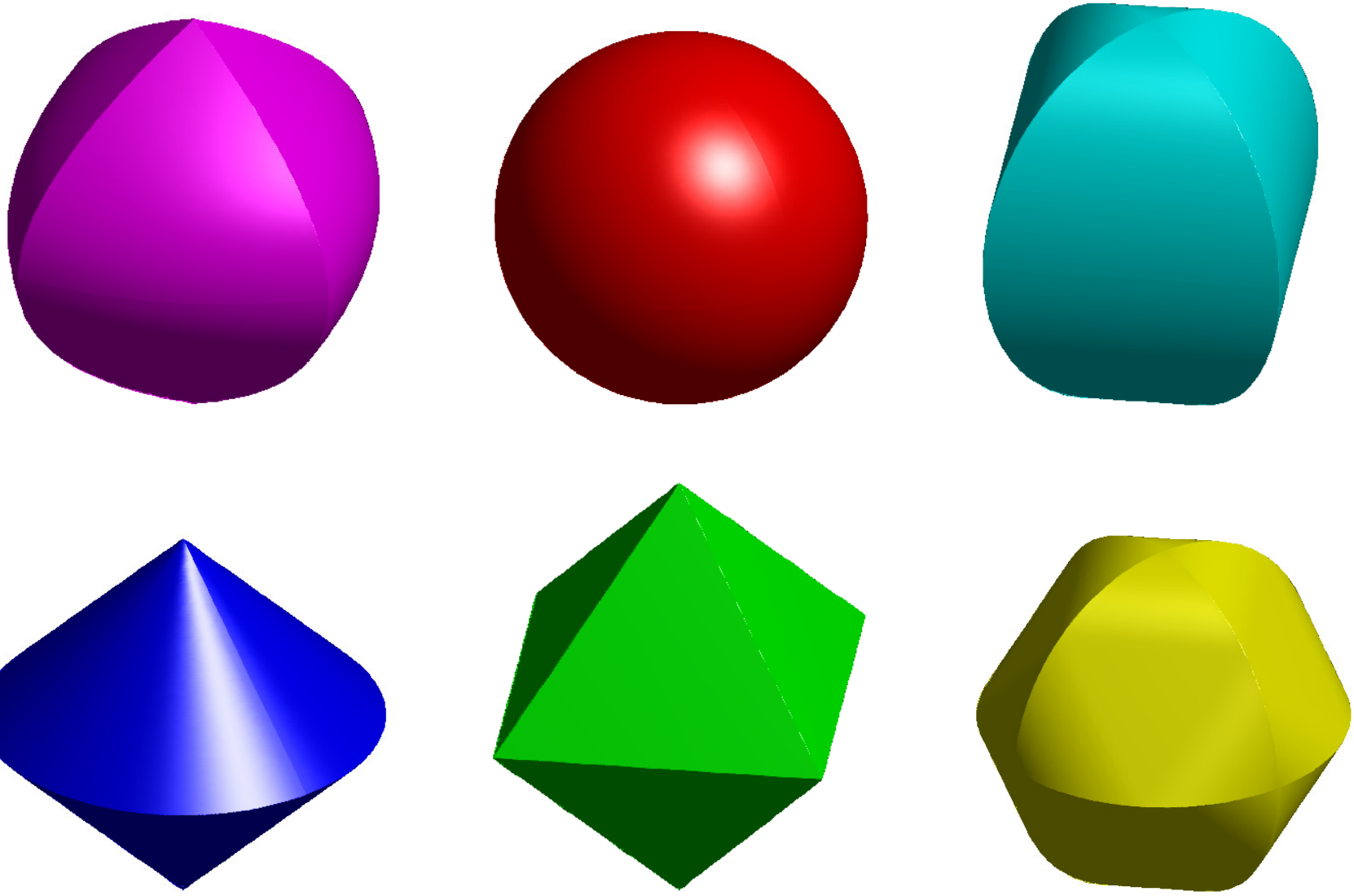
$$\|x\|^*_{\mathcal{A}} = \sup_{a\in\mathcal{A}} \langle a, x\rangle$$

Examples

- $\mathcal{A} = \{\pm e_k \mid k \in 1, \ldots, p\}$ $\quad \to \ell_1$-norm
- Let $\{G_1, \ldots, G_k\}$ be a partition of $\{1, \ldots, p\}$. Then $\mathcal{A} = \cup_i \{u \mid \mathrm{supp}(u) \subset G_i,\ \|u\|_2 = 1\}$ $\quad \to \ell_1/\ell_2$
- the trace norm is also an atomic norm

Note: the definition of the convex relaxation is *intrinsic*.

Some vectors norms



Tangent cone and Normal cone

$$\mathcal{T}_{\mathcal{A}}(x) = \text{conicHull}\{z - x \mid \|z\|_{\mathcal{A}} \leq \|x\|_{\mathcal{A}}\}$$

$$\mathcal{N}_{\mathcal{A}}(x) = \{s \mid \forall h \in \mathcal{T}_{\mathcal{A}}(x), \quad \langle s, h \rangle \leq 0\}$$

General Null Space property

Consider the optimization problem

$$\min_x \|x\|_{\mathcal{A}} \quad \text{s.t.} \quad y = \Phi x \tag{1}$$

Theorem (NSP)

x^* is the unique optimal solution of (1) if and only if

$$\mathrm{Ker}(\Phi) \cap \mathcal{T}_{\mathcal{A}}(x^*) = \varnothing.$$

Proof.

Problem (1) is equivalent to

$$\min_h \|x^* + h\|_{\mathcal{A}} \quad \text{s.t.} \quad h \in \mathrm{Ker}(\Phi).$$

If if x^* is the unique optimum, for all $h \in \mathrm{Ker}(\Phi)\backslash\{0\}$ we must have $\|x^* + h\|_{\mathcal{A}} > \|x^*\|_{\mathcal{A}}$. Conversely, $\mathrm{Ker}(\Phi) \cap \mathcal{T}_{\mathcal{A}}(x^*) = \varnothing$ implies there is not feasible descent direction from x^*. □

Note: this motivates a posteriori the construction of atomic norms.

Noisy case and minimal eigenvalue on the tangent cone

Consider the optimization problem

$$\min_x \|x\|_{\mathcal{A}} \quad \text{s.t.} \quad \|y - \Phi x\|_2 \leq \delta \tag{2}$$

Assumption (TCE(κ))

There exists $\kappa > 0$ *such that* $\quad \forall h \in \mathcal{T}_{\mathcal{A}}(x^*), \quad \|\Phi x\| \geq \kappa \|x\|.$

Theorem

Given noisy measurements $y = \Phi x^* + \epsilon$ *with* $\|\epsilon\| \leq \delta$,
if TCE(κ) *holds, then the solution* $\hat{x}$ *of (2) satisfies* $\|\hat{x} - x^*\| \leq \frac{2\delta}{\kappa}$.

Proof.

Let $\hat{x}$ be an optimal solution. We have $\|\hat{x}\|_{\mathcal{A}} \leq \|x^*\|_{\mathcal{A}}$, hence

$$\kappa \|\hat{x} - x^*\| \leq \|\Phi(\hat{x} - x^*)\| \leq \|\Phi\hat{x} - y\| + \|\Phi x^* - y\| \leq 2\delta.$$

□

Gaussian widths

Let S be a set, its Gaussian width or Gelfand width is defined as

$$w(S) = \mathbb{E}\left[\sup_{z \in S} \langle g, z \rangle\right] \quad \text{for} \quad g \sim \mathcal{N}(0, I)$$

Work of russian mathematicians in the 70ies and 80ies (Kashin, 1977; Gluskin, 1984; Garnaev and Gluskin, 1984).

Expectation of the χ distribution

Let $\lambda_n = \mathbb{E}\left[\|g\|_2\right]$ for $g \sim \mathcal{N}(0, I_n)$. We have

$$\frac{\sqrt{n}}{\sqrt{1 + \frac{1}{n}}} \leq \lambda_n \leq \sqrt{n}$$

Restricted eigenvalue for a Gaussian design

Let $\mathbb{S}^{p-1}$ denote the hypersphere in $\mathbb{R}^p$.

Theorem (Gordon (1988))

- *Let* Ω *be a closed subset of* $\mathbb{S}^{p-1}$
- *Let* $\tilde{\Phi} \in \mathbb{R}^{p \times n}$ *with i.i.d.* $\mathcal{N}(0,1)$ *entries.*

Then

$$\mathbb{E}\left[\min_{z \in \Omega} \|\tilde{\Phi} z\|_2\right] \geq \lambda_n - w(\Omega).$$

Sample complexity

The *sample complexity* is the smallest number of samples needed to guarantee that with high probability an estimator is correct or approximately correct at some level $\tilde{\delta}$.

Types of problems:

- Exact recovery $\hat{x} = x^*$
- Support recovery: $\hat{S} = S^*$
- Consistency in the ℓ_2-norm: $\|\hat{x} - x\|_2 \leq \tilde{\delta}$

Sample complexity for exact recovery in the noiseless setting

Consider again

$$\min_x \|x\|_{\mathcal{A}} \quad \text{s.t.} \quad y = \Phi x \tag{1}$$

Theorem

- *Let* $\Omega = \mathcal{T}_{\mathcal{A}}(x^*) \cap \mathbb{S}^{p-1}$
- *Let* $\Phi \in \mathbb{R}^{p \times n}$ *with i.i.d.* $\mathcal{N}(0, \frac{1}{n})$ *entries.*
- *Let* $y = \Phi x^*$

Then x^* *is the unique solution of (1) with probability at least*

$$1 - \exp\Big(-\frac{1}{2}(\lambda_n - w(\Omega))^2 \Big) \qquad \text{as soon as} \qquad n \geq w(\Omega)^2 + 1.$$

Proof

Need to show that w.h.p. $\min_{h\in\Omega} \|\Phi h\| > 0$.
But by Gordon's theorem, we know that

$$\mathbb{E}\left[\min_{h\in\Omega} \|\Phi h\|\right] \geq \frac{\lambda_n - w(\Omega)}{\sqrt{n}} > 0.$$

$$\lambda_n \geq \sqrt{\frac{n}{1+\frac{1}{n}}} \geq \sqrt{\frac{w(\Omega)^2+1}{1+\frac{1}{n}}} \geq \sqrt{\frac{w(\Omega)^2+w(\Omega)^2/n}{1+\frac{1}{n}}} \geq w(\Omega).$$

Now we know that if $g \sim \mathcal{N}(0, I_n)$ and f is a Lipschitz function with constant L then

$$\mathbb{P}\left(f(g) < \mathbb{E}\left[f(g)\right] - t\right) \leq \exp\Big(-\frac{t^2}{2L^2}\Big).$$

Since $\tilde{\Phi} \mapsto \frac{1}{\sqrt{n}} \min_{h\in\Omega} \|\tilde{\Phi} h\| = \min_z \|\Phi z\|$ is Lipschitz with constant $1/\sqrt{n}$, we thus get

$$\mathbb{P}\left(\min_{h\in\Omega} \|\Phi h\| \geq \kappa\right) \geq 1 - \exp\Big(-\frac{1}{2}(\lambda_n - w(\Omega) - \sqrt{n}\kappa)^2\Big)$$

Sample complexity for exact recovery in the noiseless setting

Consider again

$$\min_x \|x\|_{\mathcal{A}} \quad \text{s.t.} \quad y = \Phi x \tag{1}$$

Theorem

- *Let* $\Omega = \mathcal{T}_{\mathcal{A}}(x^*) \cap \mathbb{S}^{p-1}$
- *Let* $\Phi \in \mathbb{R}^{p \times n}$ *with i.i.d.* $\mathcal{N}(0, \frac{1}{n})$ *entries.*
- *Let* $y = \Phi x^*$

Then x^* *is the unique solution of (1) with probability at least*

$$1 - \exp\Big(-\frac{1}{2}(\lambda_n - w(\Omega))^2 \Big) \qquad \text{as soon as} \qquad n \geq w(\Omega)^2 + 1.$$

Sample complexity for ℓ_2 approximation in the noisy setting

Consider

$$\min_x \|x\|_{\mathcal{A}} \quad \text{s.t.} \quad \|y - \Phi x\|_2 \leq \delta \tag{2}$$

Theorem

- *Let* $\Omega = \mathcal{T}_{\mathcal{A}}(x^*) \cap \mathbb{S}^{p-1}$
- *Let* $\Phi \in \mathbb{R}^{p \times n}$ *with i.i.d.* $\mathcal{N}(0, \frac{1}{n})$ *entries.*
- *Let* $y = \Phi x^* + \epsilon$ *with* $\|\epsilon\| \leq \delta$.

Then if $\hat{x}$ *is the solution of (2) we have* $\|\hat{x} - x^*\| \leq \frac{2\delta}{\kappa}$ *with probability at least*

$$1 - \exp\left(-\frac{1}{2}(\lambda_n - w(\Omega) - \sqrt{n}\kappa)^2\right) \qquad \text{as soon as} \qquad n \geq \frac{w(\Omega)^2 + 1.5}{(1-\kappa)^2}.$$

Proof

We still have

$$\mathbb{P}\left(\min_z \|\Phi z\| \geq \kappa\right) \geq 1 - \exp\left(-\frac{1}{2}(\lambda_n - w(\Omega) - \sqrt{n}\kappa)^2\right).$$

Now under the assumptions of the theorem we have

$$w(\Omega)^2+1 \leq n(1-\kappa)^2-\frac{1}{2} \leq n(1-\kappa)^2-2\kappa(1-\kappa)+\frac{\kappa^2}{n} \leq \left(\sqrt{n}(1-\kappa)-\frac{\kappa}{\sqrt{n}}\right)^2$$

so that

$$\lambda_n-\sqrt{n}\kappa \geq \frac{n-(n+1)\kappa}{\sqrt{n+1}} = \frac{\sqrt{n}(1-\kappa)-\frac{\kappa}{\sqrt{n}}}{\sqrt{1+\frac{1}{n}}} \geq \sqrt{\frac{w(\Omega)^2+1}{1+\frac{1}{n}}} \geq w(\Omega).$$

Sample complexity for ℓ_2 approximation in the noisy setting

Consider

$$\min_x \|x\|_{\mathcal{A}} \quad \text{s.t.} \quad \|y - \Phi x\|_2 \leq \delta \tag{2}$$

Theorem

- *Let* $\Omega = \mathcal{T}_{\mathcal{A}}(x^*) \cap \mathbb{S}^{p-1}$
- *Let* $\Phi \in \mathbb{R}^{p \times n}$ *with i.i.d.* $\mathcal{N}(0, \frac{1}{n})$ *entries.*
- *Let* $y = \Phi x^* + \epsilon$ *with* $\|\epsilon\| \leq \delta$.

Then if $\hat{x}$ *is the solution of (2) we have* $\|\hat{x} - x^*\| \leq \frac{2\delta}{\kappa}$ *with probability at least*

$$1 - \exp\left(-\frac{1}{2}(\lambda_n - w(\Omega) - \sqrt{n}\kappa)^2\right) \qquad \text{as soon as} \qquad n \geq \frac{w(\Omega)^2 + 1.5}{(1-\kappa)^2}.$$

Sample complexity results from Gaussian width

ℓ_1-norm

If $x^* \in \mathbb{R}^p$ is a k sparse vector, and $\|\cdot\|_{\mathcal{A}} = \|\cdot\|_1$ then

$$w(\mathcal{T}_{\mathcal{A}}(x^*) \cap \mathbb{S}^{p-1})^2 \leq 2k \log\left(\frac{p}{k}\right) + \frac{5}{4}k,$$

and thus x^* is recovered with high probability by Basis pursuit as soon as

$$n \geq 2k \log\left(\frac{p}{k}\right) + \frac{5}{4}k + 1.$$

Trace norm

If $X^* \in \mathbb{R}^{m_1 \times m_2}$ is a rank r sparse matrix, and $\|\cdot\|_{\mathcal{A}} = \|\cdot\|_{\mathrm{tr}}$ then

$$w(\mathcal{T}_{\mathcal{A}}(X^*) \cap \mathbb{S}^{m_1 m_2 - 1})^2 \leq 3r\left(m_1 + m_2 - r\right)$$

and thus X^* is recovered with high probability by trace norm minimization as soon as $n \geq 3r\left(m_1 + m_2 - r\right) + 1$.

Take Home messages

- Greedy algorithm and convex formulations with efficient algorithms
- Support recovery/ ℓ_2-recovery require some conditions on the matrix (RIP,IC,MI,RE)
  - (For predictions no conditions for slow rate but same for fast rate)
- Realistic conditions hard to check (NP-hard)
- Random construction gives a probabilistic guarantee
- Deterministic constructions possible but for unpractically large settings as of now
- *Gaussian width* $\rightarrow$ the possibility depends on random projections of $\mathcal{T}_{\mathcal{A}}(x^*) \cap \mathbb{S}^{p-1}$.

References I

Baraniuk, R., Davenport, M., DeVore, R., and Wakin, M. (2008). A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263.

Cai, T., Wang, L., and Xu, G. (2010). Shifting inequality and recovery of sparse signals. *Signal Processing, IEEE Transactions on*, 58(3):1300–1308.

Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351.

Candes, E. J. and Tao, T. (2005). Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215.

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61.

Donoho, D. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via ?1 minimization. *Proceedings of the National Academy of Sciences of the United States of America*, 100(5):2197.

Garnaev, A. Y. and Gluskin, E. D. (1984). The widths of a euclidean ball. In *Dokl. Akad. Nauk SSSR*, volume 277, pages 1048–1052.

Gluskin, E. (1984). Norms of random matrices and widths of finite-dimensional sets. *Mathematics of the USSR-Sbornik*, 48(1):173.

References II

Gordon, Y. (1988). On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. *Geometric Aspects of Functional Analysis, Israel Seminar 1986-87, Lecture Notes in Mathematics*, 1317:84–106.

Kashin, B. S. (1977). Diameters of some finite-dimensional sets and classes of smooth functions. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, 41(2):334–351.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of The Royal Statistical Society Series B*, 58(1):267–288.

Wainwright, M. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ?1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.