A short introduction to parsimony

and

a complexity analysis of the Lasso regularization path

Julien Mairal

Inria, LEAR team, Grenoble

Toulouse, October 2014

Big data

An ill-defined concept

- a “**buzz**” word; regardless of rationality, you may get some funding and become famous by making extensive use of “big data”;
- replacing “thinking” by “data” and hope for the best;
- a means to make money from your personal data.

Big data

An ill-defined concept

- a “**buzz**” word; regardless of rationality, you may get some funding and become famous by making extensive use of “big data”;
- replacing “thinking” by “data” and hope for the best;
- a means to make money from your personal data.

A scientific utopia

- converting data into scientific knowledge;
- better understanding the world by observing it as much as we can.

Big data

An ill-defined concept

- a "**buzz**" word; regardless of rationality, you may get some funding and become famous by making extensive use of "big data";
- replacing "thinking" by "data" and hope for the best;
- a means to make money from your personal data.

A scientific utopia

- converting data into scientific knowledge;
- better understanding the world by observing it as much as we can.

in French

Mega données.

Collaborator

The analysis of the regularization path is a joint work with Bin Yu, from UC Berkeley.

Reference

J. Mairal and B. Yu. Complexity analysis of the Lasso regularization path. ICML. 2012.

Advertisement

The introduction to parsimony is based on the material of the upcoming monograph, which will be freely available on arXiv mid-october:

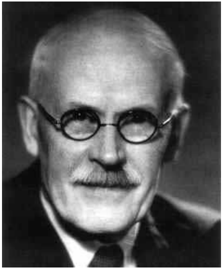
J. Mairal, F. Bach and J. Ponce. *Sparse Modeling for Image and Vision Processing*. 2014.

Part I: A Short Introduction to Parsimony

1. A short introduction to parsimony
   - Early thoughts
   - Sparsity in the statistics literature from the 60's and 70's
   - Wavelet thresholding in signal processing from 90's
   - The modern parsimony and the ℓ_1-norm
   - Structured sparsity

Early thoughts

(a) Dorothy Wrinch
1894–1980



(b) Harold Jeffreys
1891–1989

> *The existence of simple laws is, then, apparently, to be regarded as a quality of nature; and accordingly we may infer that it is justifiable to prefer a simple law to a more complex one that fits our observations slightly better.*

[Wrinch and Jeffreys, 1921]. Philosophical Magazine Series.

Historical overview of parsimony

- 14th century: Ockham's razor;
- 1921: Wrinch and Jeffreys' simplicity principle;
- 1952: Markowitz's portfolio selection;
- 60 and 70's: best subset selection in statistics;
- 70's: use of the ℓ_1-norm for signal recovery in geophysics;
- 90's: wavelet thresholding in signal processing;
- 1996: Olshausen and Field's dictionary learning;
- 1996–1999: Lasso (statistics) and basis pursuit (signal processing);
- 2006–now: compressed sensing (signal processing) and Lasso consistency (statistics); applications in various scientific fields such as image processing, bioinformatics, neuroscience, computer vision...

Sparsity in the statistics literature from the 60's and 70's

Given some observed data points $\mathbf{z}_1, \ldots, \mathbf{z}_n$ that are assumed to be independent samples from a statistical model with parameters $\boldsymbol{\theta}$ in $\mathbb{R}^p$, **maximum likelihood estimation (MLE)** consists of minimizing

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left[\mathcal{L}(\boldsymbol{\theta}) \overset{\triangle}{=} - \sum_{i=1}^{n} \log P_{\boldsymbol{\theta}}(\mathbf{z}_i) \right].$$

Example: ordinary least square

Observations $\mathbf{z}_i = (y_i, \mathbf{x}_i)$, with y_i in $\mathbb{R}$.
Linear model: $y_i = \mathbf{x}_i^\top \boldsymbol{\theta} + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0,1)$.

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \sum_{i=1}^{n} \frac{1}{2} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\theta} \right)^2.$$

Sparsity in the statistics literature from the 60's and 70's

Given some observed data points $\mathbf{z}_1, \ldots, \mathbf{z}_n$ that are assumed to be independent samples from a statistical model with parameters $\boldsymbol{\theta}$ in $\mathbb{R}^p$, **maximum likelihood estimation (MLE)** consists of minimizing

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left[\mathcal{L}(\boldsymbol{\theta}) \overset{\triangle}{=} - \sum_{i=1}^{n} \log P_{\boldsymbol{\theta}}(\mathbf{z}_i) \right].$$

Motivation for finding a sparse solution:

- removing irrelevant variables from the model;
- obtaining an easier interpretation;
- preventing overfitting;

Sparsity in the statistics literature from the 60's and 70's

Given some observed data points $\mathbf{z}_1, \ldots, \mathbf{z}_n$ that are assumed to be independent samples from a statistical model with parameters $\boldsymbol{\theta}$ in $\mathbb{R}^p$, **maximum likelihood estimation (MLE)** consists of minimizing

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left[\mathcal{L}(\boldsymbol{\theta}) \stackrel{\triangle}{=} - \sum_{i=1}^{n} \log P_{\boldsymbol{\theta}}(\mathbf{z}_i) \right].$$

Why this is highly relevant in a modern big data context:

- large n allows learning (better) complex models with large p;
- large p leads to poor interpretation and irrelevant variables;
- large p and large n lead to high computational cost;

Sparsity in the statistics literature from the 60's and 70's

Given some observed data points $\mathbf{z}_1, \ldots, \mathbf{z}_n$ that are assumed to be independent samples from a statistical model with parameters $\boldsymbol{\theta}$ in $\mathbb{R}^p$, **maximum likelihood estimation (MLE)** consists of minimizing

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left[\mathcal{L}(\boldsymbol{\theta}) \triangleq - \sum_{i=1}^{n} \log P_{\boldsymbol{\theta}}(\mathbf{z}_i) \right].$$

Two questions:

1. how to choose k?
2. how to find the best subset of k variables?

Sparsity in the statistics literature from the 60's and 70's

How to choose k?

- Mallows's C_p statistics [Mallows, 1964, 1966];
- Akaike information criterion (AIC) [Akaike, 1973];
- Bayesian information critertion (BIC) [Schwarz, 1978];
- Minimum description length (MDL) [Rissanen, 1978].

These approaches lead to penalized problems

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_0,$$

with different choices of λ depending on the chosen criterion.

Sparsity in the statistics literature from the 60's and 70's

How to solve the best k-subset selection problem?

Unfortunately...

...the problem is NP-hard [Natarajan, 1995].

Two strategies

- **combinatorial exploration** with branch-and-bound techniques [Furnival and Wilson, 1974] → **leaps and bounds**, exact algorithm but exponential complexity;
- **greedy approach**: forward selection [Efroymson, 1960] (originally developed for observing *intermediate* solutions), already contains all the ideas of **matching pursuit** algorithms.

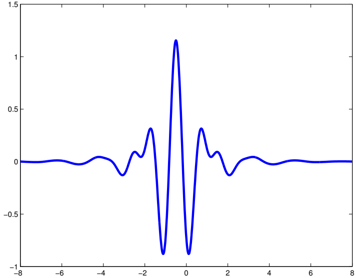
Important reference: [Hocking, 1976]. *The analysis and selection of variables in linear regression*. Biometrics.

Wavelet thresholding in signal processing from the 90's

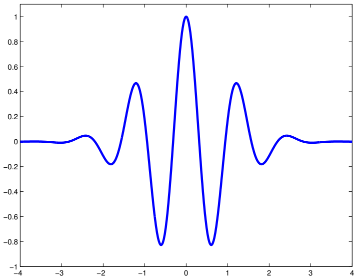
A wavelet basis represents a set of functions φ_1, φ_2 that are essentially dilated and shifted versions of each other [see Mallat, 2008].

Concept of parsimony with wavelets

When a signal f is "smooth", it is close to an expansion $\sum_i \alpha_i \varphi_i$ where only a few coefficients α_i are non-zero.



(a) Meyer's wavelet

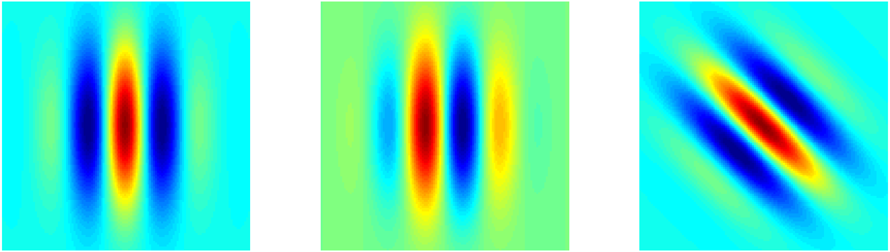


(b) Morlet's wavelet

Wavelet thresholding in signal processing from the 90's

Wavelets where the topic of a long quest for representing natural images

- 2D-Gabors [Daugman, 1985];
- steerable wavelets [Simoncelli et al., 1992];
- curvelets [Candès and Donoho, 2002];
- countourlets [Do and Vertterli, 2003];
- bandlets [Le Pennec and Mallat, 2005];
- $\star$-lets.



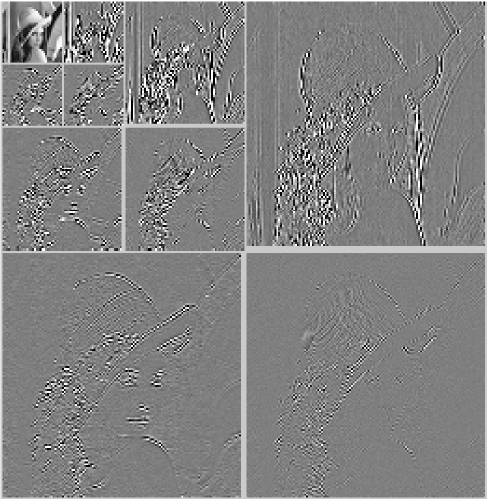
(a) 2D Gabor filter.

(b) With shifted phase.

(c) With rotation.

Wavelet thresholding in signal processing from 90's

The theory of wavelets is well developed for continuous signals, *e.g.*, in $L^2(\mathbb{R})$, but also for discrete signals $\mathbf{x}$ in $\mathbb{R}^n$.



Wavelet thresholding in signal processing from 90's

Given an orthogonal wavelet basis $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_n]$ in $\mathbb{R}^{n \times n}$, the wavelet decomposition of $\mathbf{x}$ in $\mathbb{R}^n$ is simply

$$\boldsymbol{\beta} = \mathbf{D}^\top \mathbf{x} \quad \text{and we have} \quad \mathbf{x} = \mathbf{D}\boldsymbol{\beta}.$$

The k-sparse approximation problem

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\alpha}\|_0 \leq k,$$

is not NP-hard here: since $\mathbf{D}$ is orthogonal, it is equivalent to

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2} \|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\alpha}\|_0 \leq k.$$

Wavelet thresholding in signal processing from 90's

Given an orthogonal wavelet basis $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_n]$ in $\mathbb{R}^{n \times n}$, the wavelet decomposition of $\mathbf{x}$ in $\mathbb{R}^n$ is simply

$$\boldsymbol{\beta} = \mathbf{D}^\top \mathbf{x} \quad \text{and we have} \quad \mathbf{x} = \mathbf{D}\boldsymbol{\beta}.$$

The k-sparse approximation problem

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\alpha}\|_0 \leq k,$$

The solution is obtained by **hard-thresholding**:

$$\boldsymbol{\alpha}^{\text{ht}}[j] = \delta_{|\boldsymbol{\beta}[j]| \geq \mu} \boldsymbol{\beta}[j] = \begin{cases} \boldsymbol{\beta}[j] & \text{if } \ |\boldsymbol{\beta}[j]| \geq \mu \\ 0 & \qquad \text{otherwise} \end{cases},$$

where μ the k-th largest value among the set $\{|\boldsymbol{\beta}[1]|, \ldots, |\boldsymbol{\beta}[p]|\}$.

Wavelet thresholding in signal processing, 90's

Another key operator introduced by Donoho and Johnstone [1994] is the **soft-thresholding** operator:

$$\boldsymbol{\alpha}^{\text{st}}[j] \triangleq \text{sign}(\beta[j]) \max(|\beta[j]| - \lambda, 0) = \begin{cases} \beta[j] - \lambda & \text{if } \beta[j] \geq \lambda \\ \beta[j] + \lambda & \text{if } \beta[j] \leq -\lambda \\ 0 & \text{otherwise} \end{cases},$$

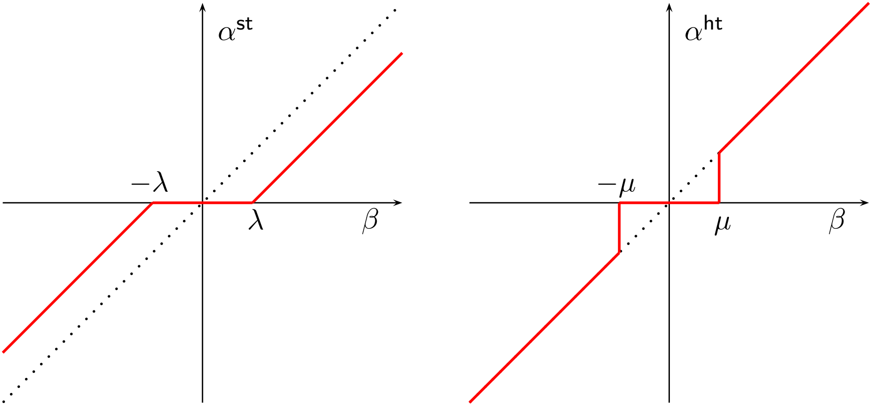
where λ is a parameter playing the same role as μ previously.

With $\boldsymbol{\beta} \triangleq \mathbf{D}^\top \mathbf{x}$ and $\mathbf{D}$ orthogonal, it provides the solution of the following sparse reconstruction problem:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1,$$

which will be of high importance later.

Wavelet thresholding in signal processing, 90's



(d) Soft-thresholding operator, $\alpha^{\text{st}} = \text{sign}(\beta)\max(|\beta| - \lambda, 0)$.

(e) Hard-thresholding operator $\alpha^{\text{ht}} = \delta_{|\beta| \geq \mu}\beta$.

Figure : Soft- and hard-thresholding operators, which are commonly used for signal estimation with orthogonal wavelet basis.

Wavelet thresholding in signal processing, 90's

Various work tried to exploit the **structure** of wavelet coefficients.

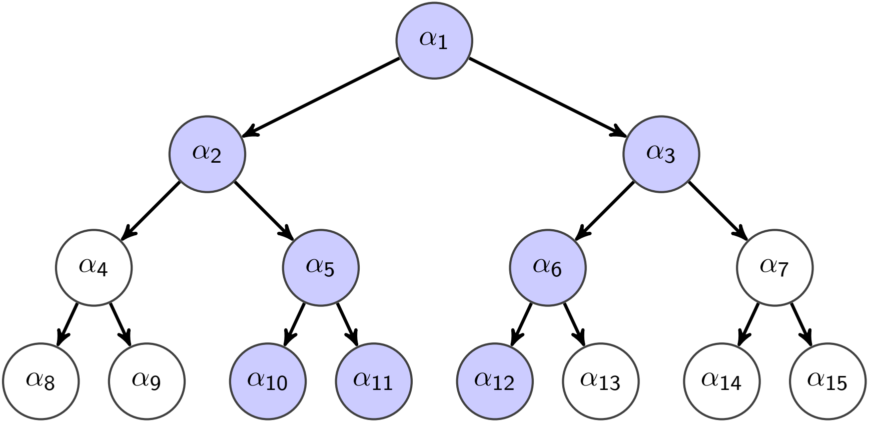
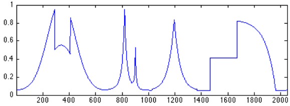


Figure : Illustration of a wavelet tree with four scales for one-dimensional signals. We also illustrate the zero-tree coding scheme [Shapiro, 1993].

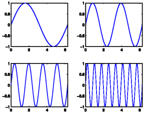
The modern parsimony and the ℓ_1-norm

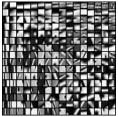
Sparse linear models in signal processing

Let $\mathbf{x}$ in $\mathbb{R}^n$ be a signal.



Let $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_p] \in \mathbb{R}^{n \times p}$ be a set of elementary signals.
We call it **dictionary**.





$\mathbf{D}$ is "adapted" to $\mathbf{x}$ if it can represent it with a few elements—that is, there exists a **sparse vector** $\boldsymbol{\alpha}$ in $\mathbb{R}^p$ such that $\mathbf{x} \approx \mathbf{D}\boldsymbol{\alpha}$. We call $\boldsymbol{\alpha}$ the **sparse code**.

$$\underbrace{\begin{pmatrix} \mathbf{x} \end{pmatrix}}_{\mathbf{x}\in\mathbb{R}^n} \approx \underbrace{\left(\begin{array}{c|c|c|c} \mathbf{d}_1 & \mathbf{d}_2 & \cdots & \mathbf{d}_p \end{array}\right)}_{\mathbf{D}\in\mathbb{R}^{n\times p}} \underbrace{\begin{pmatrix} \boldsymbol{\alpha}[1] \\ \boldsymbol{\alpha}[2] \\ \vdots \\ \boldsymbol{\alpha}[p] \end{pmatrix}}_{\boldsymbol{\alpha}\in\mathbb{R}^p, \textbf{sparse}}$$

The modern parsimony and the ℓ_1-norm

Sparse linear models: machine learning/statistics point of view

Let $(y_i, \mathbf{x}_i)_{i=1}^n$ be a training set, where the vectors $\mathbf{x}_i$ are in $\mathbb{R}^p$ and are called features. The scalars y_i are in

- $\{-1, +1\}$ for **binary** classification problems.
- $\mathbb{R}$ for **regression** problems.

We assume there exists a relation $y \approx \boldsymbol{\beta}^\top \mathbf{x}$, and solve

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \underbrace{\frac{1}{n} \sum_{i=1}^{n} L(y_i, \boldsymbol{\beta}^\top \mathbf{x}_i)}_{\text{empirical risk}} + \underbrace{\lambda \psi(\boldsymbol{\beta})}_{\text{regularization}} .$$

The modern parsimony and the ℓ_1-norm

Sparse linear models: machine learning/statistics point of view

A few examples:

Ridge regression:

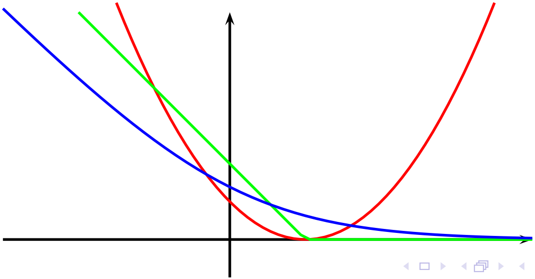
$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2 + \lambda\|\boldsymbol{\beta}\|_2^2.$$

Linear SVM:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i\boldsymbol{\beta}^\top \mathbf{x}_i) + \lambda\|\boldsymbol{\beta}\|_2^2.$$

Logistic regression:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y_i\boldsymbol{\beta}^\top \mathbf{x}_i}\right) + \lambda\|\boldsymbol{\beta}\|_2^2.$$



The modern parsimony and the ℓ_1-norm

Sparse linear models: machine learning/statistics point of view

A few examples:

Ridge regression:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}_p} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2 + \lambda\|\boldsymbol{\beta}\|_2^2.$$

Linear SVM:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}_p} \frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - y_i\boldsymbol{\beta}^\top \mathbf{x}_i) + \lambda\|\boldsymbol{\beta}\|_2^2.$$

Logistic regression:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}_p} \frac{1}{n}\sum_{i=1}^{n}\log\left(1 + e^{-y_i\boldsymbol{\beta}^\top \mathbf{x}_i}\right) + \lambda\|\boldsymbol{\beta}\|_2^2.$$

The **squared ℓ_2-norm** induces "**smoothness**" in $\boldsymbol{\beta}$. When one knows in advance that $\boldsymbol{\beta}$ should be sparse, one should use a **sparsity-inducing** regularization such as the **ℓ_1-norm**. [Chen et al., 1999, Tibshirani, 1996]

The modern parsimony and the ℓ_1-norm

Originally used to induce sparsity in geophysics [Claerbout and Muir, 1973, Taylor et al., 1979], the ℓ_1-norm became popular in statistics with the **Lasso** [Tibshirani, 1996] and in signal processing with the **Basis pursuit** [Chen et al., 1999].

Three "equivalent" formulations

1. $$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1;$$

2. $$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\alpha}\|_1 \leq \mu;$$

3. $$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\boldsymbol{\alpha}\|_1 \ \text{ s.t. } \ \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \leq \varepsilon.$$

The modern parsimony and the ℓ_1-norm

And some variants...

For noiseless problems

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\boldsymbol{\alpha}\|_1 \ \text{ s.t. } \ \mathbf{x} = \mathbf{D}\boldsymbol{\alpha}.$$

Beyond least squares

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} f(\boldsymbol{\alpha}) + \lambda \|\boldsymbol{\alpha}\|_1,$$

where $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is convex.

The modern parsimony and the ℓ_1-norm

And some variants...

For noiseless problems

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\boldsymbol{\alpha}\|_1 \ \text{ s.t. } \ \mathbf{x} = \mathbf{D}\boldsymbol{\alpha}.$$

Beyond least squares

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} f(\boldsymbol{\alpha}) + \lambda \|\boldsymbol{\alpha}\|_1,$$

where $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is convex.

An important question remains:

why does the ℓ_1-norm induce sparsity?

The modern parsimony and the ℓ_1-norm

Why does the ℓ_1-norm induce sparsity?

Can we get some intuition from the simplest isotropic case?

$$\hat{\boldsymbol{\alpha}}(\lambda) = \operatorname*{arg\,min}_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1,$$

or equivalently the Euclidean projection onto the ℓ_1-ball?

$$\tilde{\boldsymbol{\alpha}}(\mu) = \operatorname*{arg\,min}_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \boldsymbol{\alpha}\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\alpha}\|_1 \leq \mu.$$

"equivalent" means that for all $\lambda \geq 0$, there exists $\mu \geq 0$ such that $\tilde{\boldsymbol{\alpha}}(\mu) = \hat{\boldsymbol{\alpha}}(\lambda)$ and vice versa.

The modern parsimony and the ℓ_1-norm

Why does the ℓ_1-norm induce sparsity?

Can we get some intuition from the simplest isotropic case?

$$\hat{\boldsymbol{\alpha}}(\lambda) = \operatorname*{arg\,min}_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1,$$

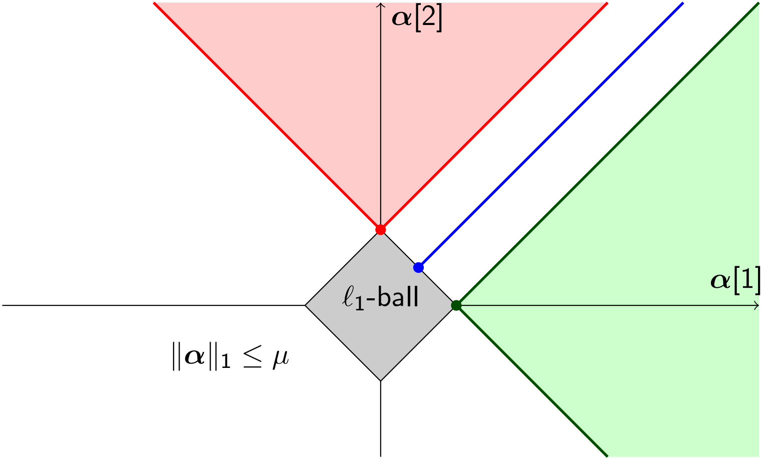
or equivalently the Euclidean projection onto the ℓ_1-ball?

$$\tilde{\boldsymbol{\alpha}}(\mu) = \operatorname*{arg\,min}_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \boldsymbol{\alpha}\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\alpha}\|_1 \leq \mu.$$

"equivalent" means that for all $\lambda \geq 0$, there exists $\mu \geq 0$ such that $\tilde{\boldsymbol{\alpha}}(\mu) = \hat{\boldsymbol{\alpha}}(\lambda)$ and vice versa.
The relation between μ and λ is unknown a priori.

Why does the ℓ_1-norm induce sparsity?

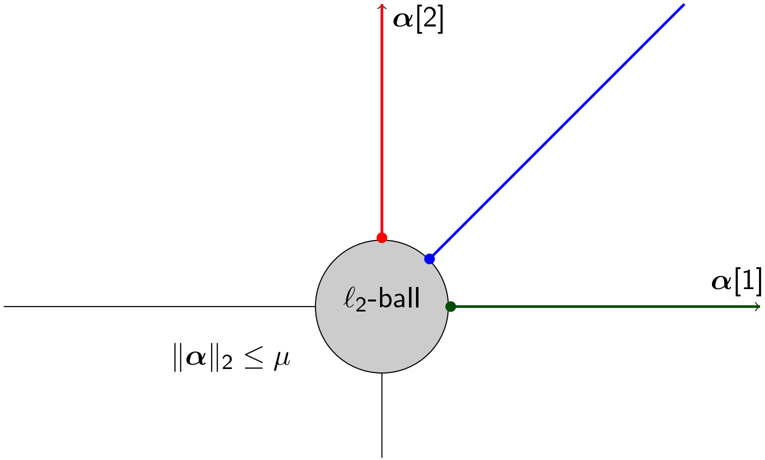
Regularizing with the ℓ_1-norm



The projection onto a convex set is “biased” towards singularities.

Why does the ℓ_1-norm induce sparsity?

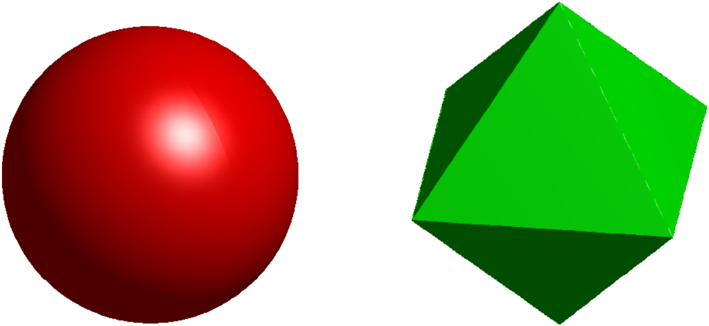
Regularizing with the ℓ_2-norm



The ℓ_2-norm is isotropic.

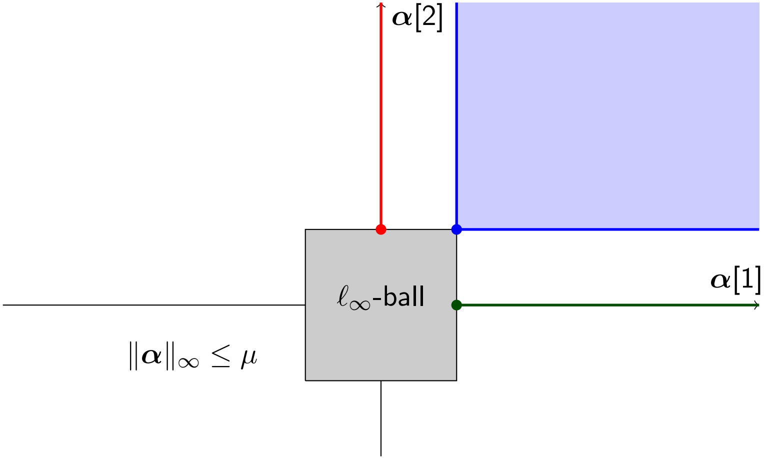
Why does the ℓ_1-norm induce sparsity?

In 3D. (images produced by G. Obozinski



Why does the ℓ_1-norm induce sparsity?

Regularizing with the ℓ_∞-norm



The ℓ_∞-norm encourages $|\boldsymbol{\alpha}[1]| = |\boldsymbol{\alpha}[2]|$.

Why does the ℓ_1-norm induce sparsity?

Analytical point of view: 1D case

$$\min_{\alpha \in \mathbb{R}} \frac{1}{2}(x-\alpha)^2 + \lambda|\alpha|$$

Piecewise quadratic function with a kink at zero.

Derivative at 0_+: $g_+ = -x + \lambda$ and 0_-: $g_- = -x - \lambda$.

Optimality conditions. α is optimal iff:

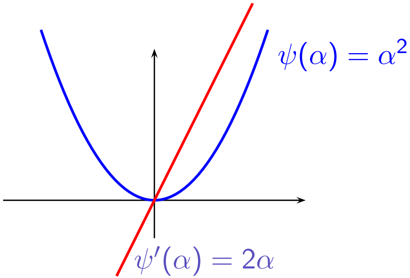
- $|\alpha| > 0$ and $(x - \alpha) + \lambda \operatorname{sign}(\alpha) = 0$
- $\alpha = 0$ and $g_+ \geq 0$ and $g_- \leq 0$

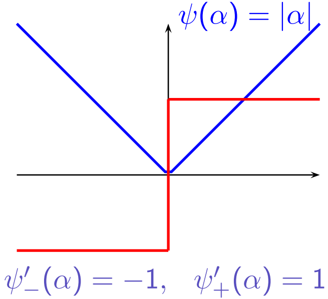
The solution is a **soft-thresholding**:

$$\alpha^\star = \operatorname{sign}(x)(|x| - \lambda)^+.$$

Why does the ℓ_1-norm induce sparsity?

Comparison with ℓ_2-regularization in 1D

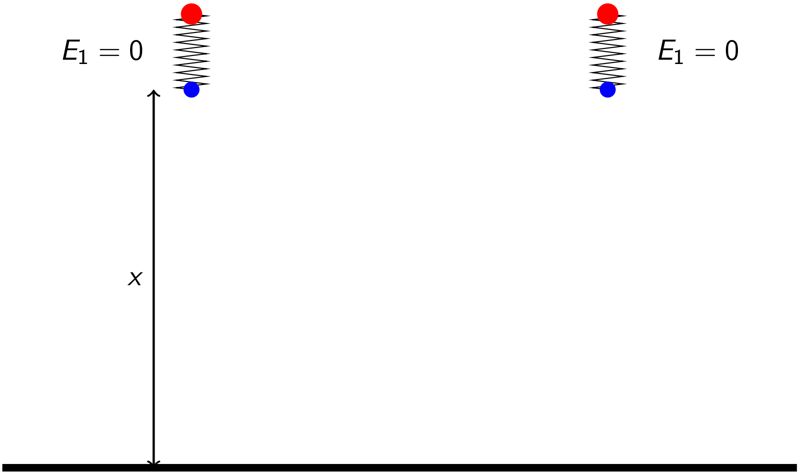




The gradient of the ℓ_2-penalty vanishes when α get close to 0. On its differentiable part, the norm of the gradient of the ℓ_1-norm is constant.

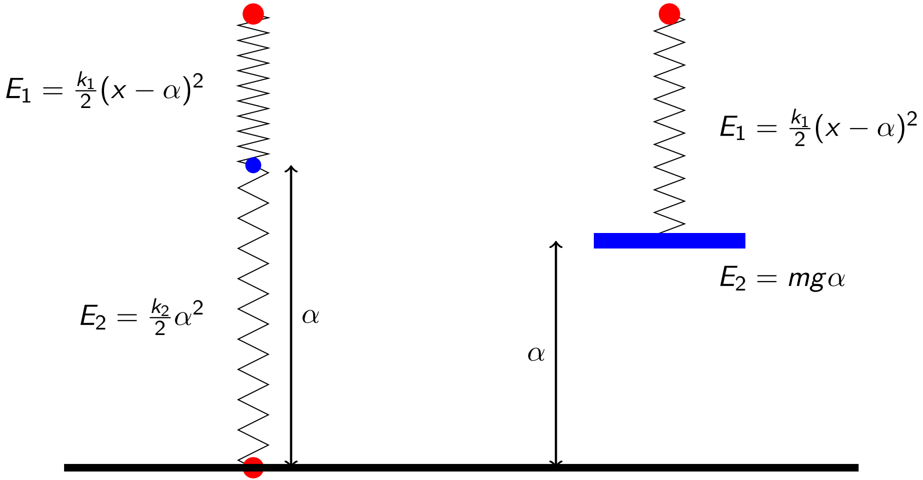
Why does the ℓ_1-norm induce sparsity?

Physical illustration



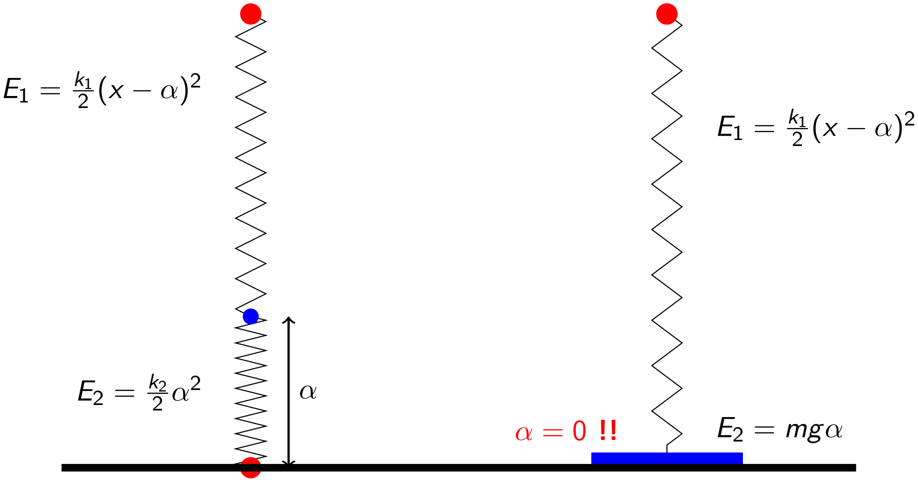
Why does the ℓ_1-norm induce sparsity?

Physical illustration


$E_1 = \frac{k_1}{2}(x-\alpha)^2$
$E_2 = \frac{k_2}{2}\alpha^2$
α
α
$E_1 = \frac{k_1}{2}(x-\alpha)^2$
$E_2 = mg\alpha$

Why does the ℓ_1-norm induce sparsity?

Physical illustration



$E_1 = \frac{k_1}{2}(x-\alpha)^2$

$E_2 = \frac{k_2}{2}\alpha^2$

α

$E_1 = \frac{k_1}{2}(x-\alpha)^2$

$\alpha = 0$ **!!**

$E_2 = mg\alpha$

Why does the ℓ_1-norm induce sparsity?

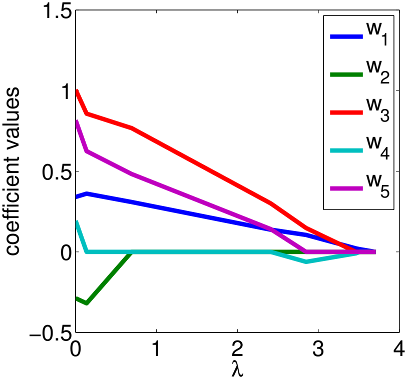


Figure : The regularization path of the Lasso.

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1.$$

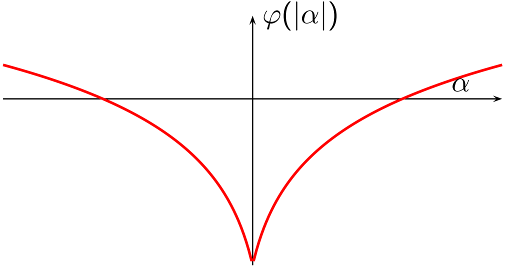
Non-convex sparsity-inducing penalties

Exploiting concave functions with a kink at zero

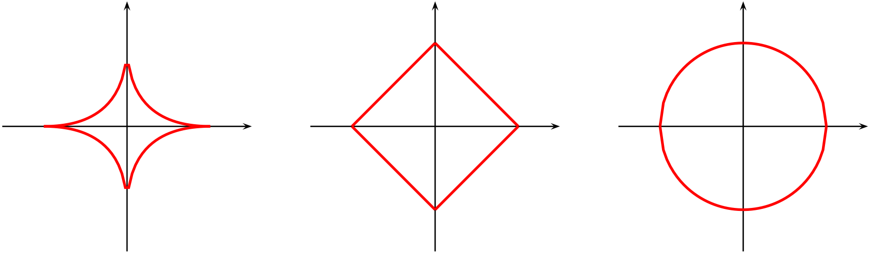
$\psi(\boldsymbol{\alpha}) = \sum_{j=1}^{p} \varphi(|\boldsymbol{\alpha}[j]|)$.

- ℓ_q-penalty, with $0 < q < 1$: $\psi(\boldsymbol{\alpha}) \triangleq \sum_{j=1}^{p} |\boldsymbol{\alpha}[j]|^q$, [Frank and Friedman, 1993];
- log penalty, $\psi(\boldsymbol{\alpha}) \triangleq \sum_{j=1}^{p} \log(|\boldsymbol{\alpha}[j]| + \varepsilon)$, [Candès et al., 2008].

φ is any function that looks like this:



Non-convex sparsity-inducing penalties



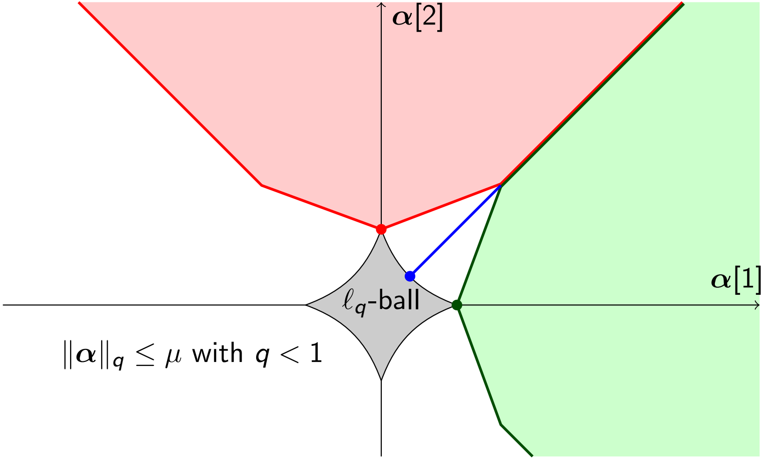
(a) $\ell_{0.5}$-ball, 2-D

(b) ℓ_1-ball, 2-D

(c) ℓ_2-ball, 2-D

Figure : Open balls in 2-D corresponding to several ℓ_q-norms and pseudo-norms.

Non-convex sparsity-inducing penalties

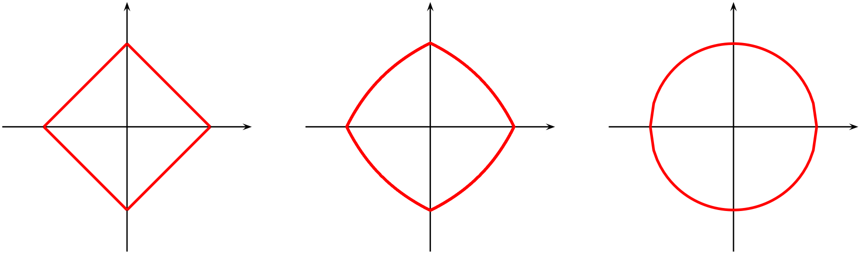


Elastic-net

The **elastic net** introduced by [Zou and Hastie, 2005]

$$\psi(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1 + \gamma \|\boldsymbol{\alpha}\|_2^2,$$

The penalty provides more stable (but less sparse) solutions.



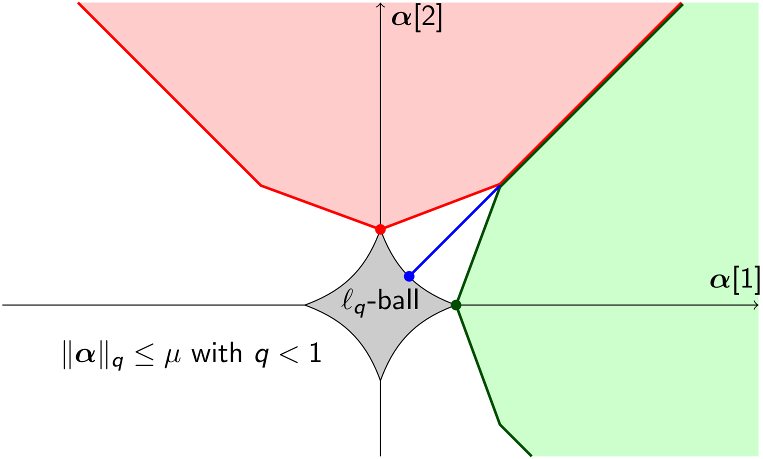
(a) ℓ_1-ball, 2-D

(b) elastic-net, 2-D

(c) ℓ_2-ball, 2-D

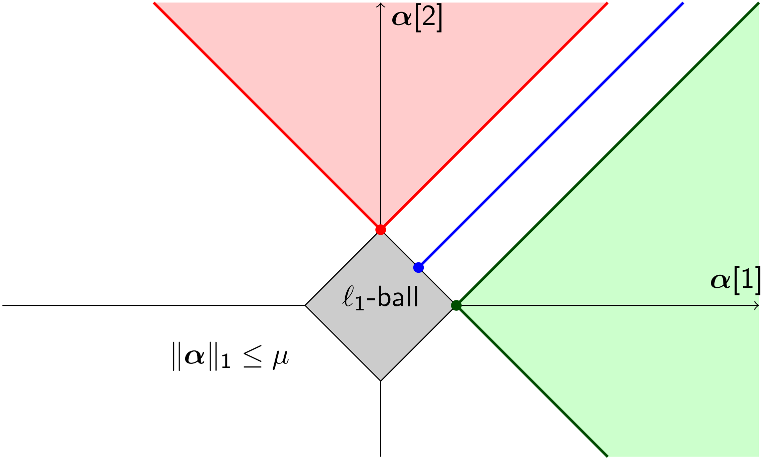
The elastic-net

vs other penalties



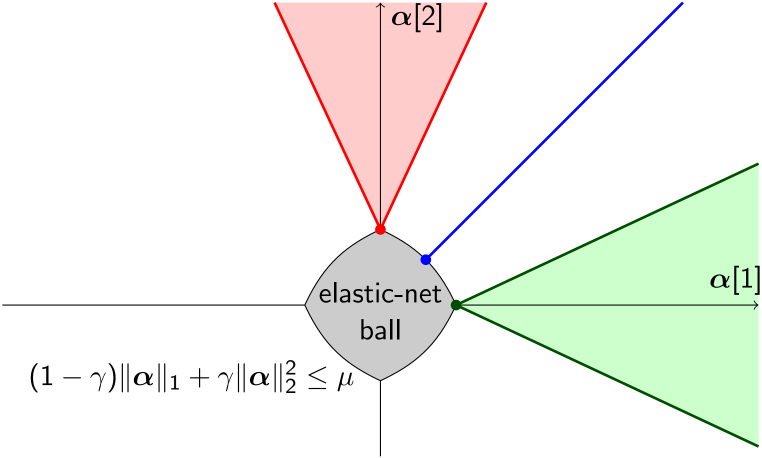
The elastic-net

vs other penalties



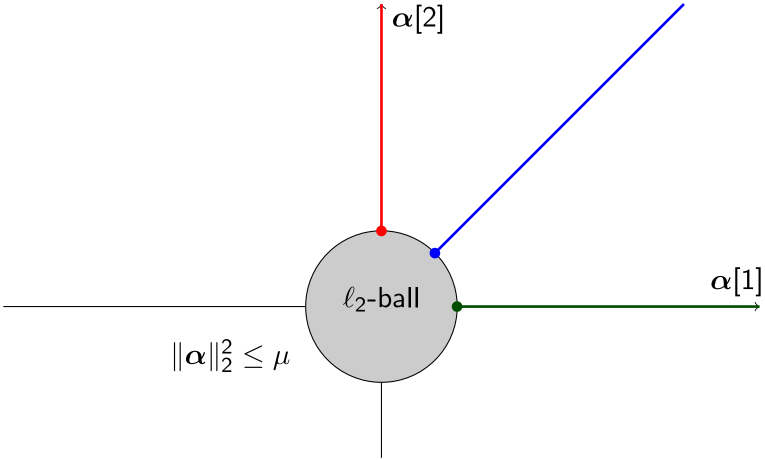
The elastic-net

vs other penalties


$\boldsymbol{\alpha}[2]$
$\boldsymbol{\alpha}[1]$
elastic-net ball
$(1-\gamma)\|\boldsymbol{\alpha}\|_1 + \gamma\|\boldsymbol{\alpha}\|_2^2 \leq \mu$

The elastic-net

vs other penalties



Total variation and fused Lasso

The anisotropic **total variation** [Rudin et al., 1992]

$$\psi(\boldsymbol{\alpha}) = \sum_{j=1}^{p-1} |\boldsymbol{\alpha}[j+1] - \boldsymbol{\alpha}[j]|,$$

called **fused Lasso** in statistics [Tibshirani et al., 2005]. The penalty encourages piecewise constant signals (can be extended to images).

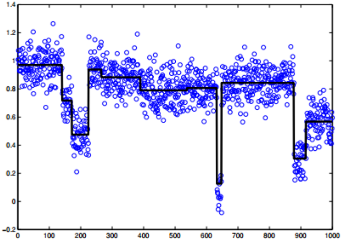


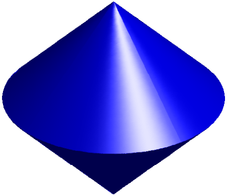
Image borrowed from a talk of J.-P. Vert, representing DNA copy numbers.

Group Lasso and mixed norms

[Turlach et al., 2005, Yuan and Lin, 2006, Zhao et al., 2009]
[Grandvalet and Canu, 1999, Bakin, 1999]

$$\text{the } \ell_1/\ell_q\text{-norm :} \qquad \psi(\boldsymbol{\alpha}) = \sum_{g \in \mathcal{G}} \|\boldsymbol{\alpha}[g]\|_q.$$

- $\mathcal{G}$ is a **partition** of $\{1, \ldots, p\}$;
- $q = 2$ or $q = \infty$ in practice;
- can be interpreted as the ℓ_1-norm of $[\|\boldsymbol{\alpha}[g]\|_q]_{g \in \mathcal{G}}$.



$$\psi(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}[\{1,2\}]\|_2 + |\boldsymbol{\alpha}[3]|.$$

Spectral sparsity

[Fazel et al., 2001, Srebro et al., 2005]

A natural regularization function for matrices is the **rank**

$$\text{rank}(\mathbf{A}) \triangleq |\{j : s_j(\mathbf{A}) \neq 0\}| = \|\mathbf{s}(\mathbf{A})\|_0,$$

where s_j is the j-th singular value and $\mathbf{s}$ is the *spectrum* of $\mathbf{A}$.

A successful convex relaxation of the rank is the sum of singular values

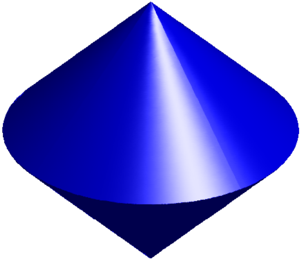
$$\|\mathbf{A}\|_* \triangleq \sum_{j=1}^{p} s_j(\mathbf{A}) = \|\mathbf{s}(\mathbf{A})\|_1,$$

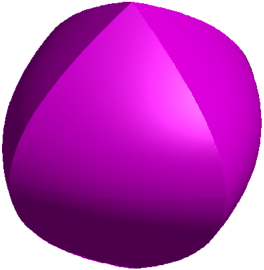
for $\mathbf{A}$ in $\mathbb{R}^{p \times k}$ with $k \geq p$.

The resulting function is a norm, called the **trace** or **nuclear** norm.

Structured sparsity

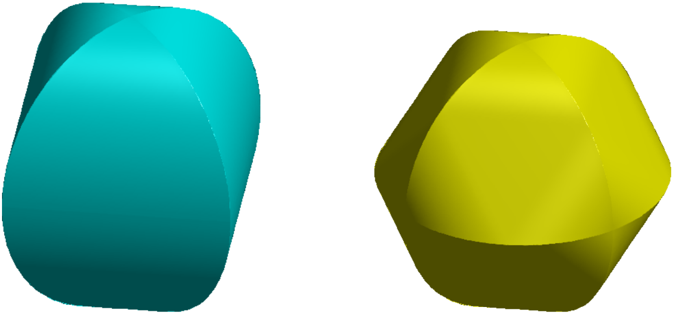
images produced by G. Obozinski





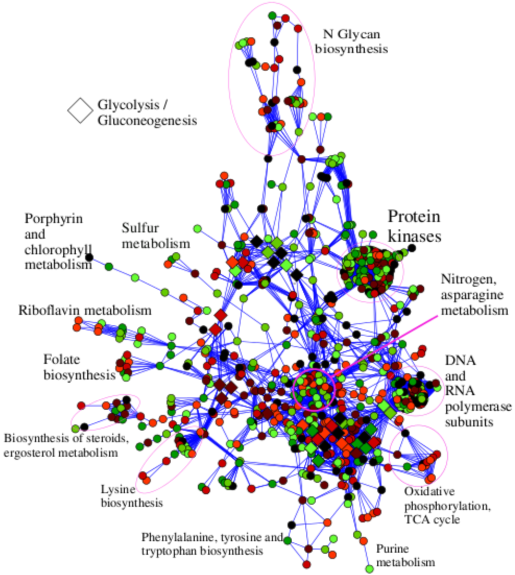
Structured sparsity

images produced by G. Obozinski



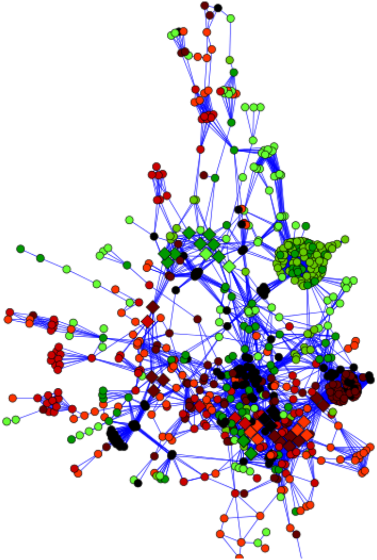
Structured sparsity

Metabolic network of the budding yeast from Rapaport et al. [2007]


N Glycan biosynthesis
◇ Glycolysis / Gluconeogenesis
Porphyrin and chlorophyll metabolism
Sulfur metabolism
Protein kinases
Riboflavin metabolism
Nitrogen, asparagine metabolism
Folate biosynthesis
DNA and RNA polymerase subunits
Biosynthesis of steroids, ergosterol metabolism
Lysine biosynthesis
Oxidative phosphorylation, TCA cycle
Phenylalanine, tyrosine and tryptophan biosynthesis
Purine metabolism

Structured sparsity

Metabolic network of the budding yeast from Rapaport et al. [2007]



Structured sparsity

Warning: Under the name "structured sparsity" appear in fact significantly different formulations!

1. non-convex
    - zero-tree wavelets [Shapiro, 1993];
    - predefined collection of sparsity patterns: [Baraniuk et al., 2010];
    - **select a union of groups: [Huang et al., 2009]**;
    - structure via Markov random fields: [Cehver et al., 2008];
2. convex (norms)
    - **tree-structure: [Zhao et al., 2009]**;
    - **select a union of groups: [Jacob et al., 2009]**;
    - **zero-pattern is a union of groups: [Jenatton et al., 2011]**;
    - other norms: [Micchelli et al., 2013].

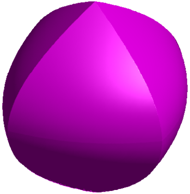
Structured sparsity

Group Lasso with overlapping groups [Jenatton et al., 2011]

$$\psi(\boldsymbol{\alpha}) = \sum_{g \in \mathcal{G}} \|\boldsymbol{\alpha}[g]\|_q .$$

What happens when the groups overlap?

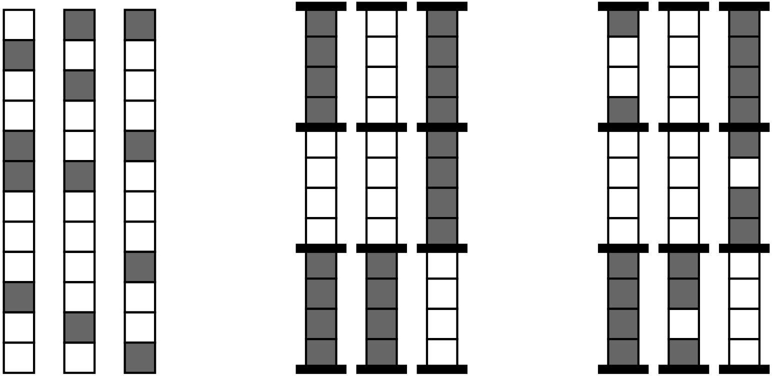
- the pattern of non-zero variables is an intersection of groups;
- the **zero pattern is a union of groups**.



$$\psi(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_2 + |\boldsymbol{\alpha}[2]| + |\boldsymbol{\alpha}[3]|.$$

Structured sparsity

Hierarchical norms [Zhao et al., 2009].



(d) Sparsity.

(e) Group sparsity.

(f) Hierarchical sparsity.

Some thoughts from Hocking [1976]:

> **The problem of selecting a subset of independent or predictor variables is usually described in an idealized setting**. *That is, it is assumed that (a) the analyst has data on a large number of potential variables which include all relevant variables and appropriate functions of them plus, possibly, some other extraneous variables and variable functions and (b) the analyst has available "good" data on which to base the eventual conclusions.* **In practice, the lack of satisfaction of these assumptions may make a detailed subset selection analysis a meaningless exercise.**

Part II: Complexity Analysis of the Lasso Regularization Path

What this work is about

- another paper about the Lasso/Basis Pursuit [Tibshirani, 1996, Chen et al., 1999]:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1; \tag{1}$$

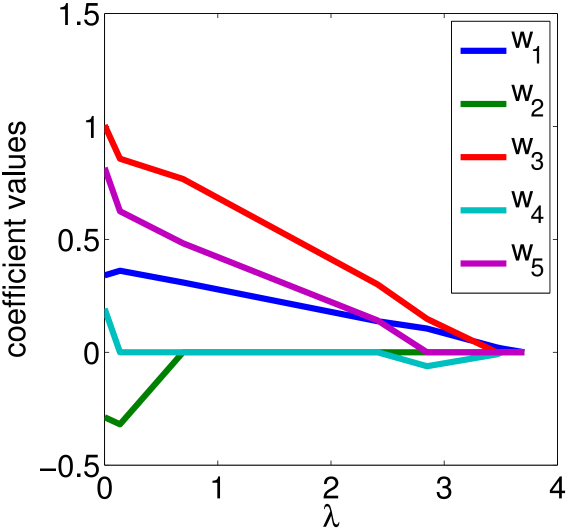
- the first complexity analysis of the homotopy method [Ritter, 1962, Osborne et al., 2000, Efron et al., 2004] for solving (1);
- a robust homotopy algorithm.

A main message reminiscent of

- the simplex algorithm [Klee and Minty, 1972];
- the SVM regularization path [Gärtner et al., 2010].

The Lasso Regularization Path and the Homotopy

When it exists, the regularization path is piecewise linear:



Our Main Results

Theorem - worst case analysis

In the worst-case, the regularization path of the Lasso has exactly $(3^p + 1)/2$ *linear segments.*

Proposition - approximate analysis

there exists an ε*-approximate path with* $O(1/\sqrt{\varepsilon})$ *linear segments.*

Brief Introduction to the Homotopy Algorithm

Optimality conditions of the Lasso

$\mathbf{w}^\star$ in $\mathbb{R}^p$ is a solution of Eq. (1) if and only if for all j in $\{1, \dots, p\}$,

$$\begin{aligned} \mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\mathbf{w}^\star) &= \lambda \, \mathrm{sign}(\mathbf{w}_j^\star) \;\; \text{if} \;\; \mathbf{w}_j^\star \neq 0, \\ |\mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\mathbf{w}^\star)| &\leq \lambda \;\; \text{otherwise.} \end{aligned}$$

Brief Introduction to the Homotopy Algorithm

Optimality conditions of the Lasso

$\mathbf{w}^\star$ in $\mathbb{R}^p$ is a solution of Eq. (1) if and only if for all j in $\{1,\ldots,p\}$,

$$\begin{aligned} \mathbf{x}^{j\top}(\mathbf{y}-\mathbf{X}\mathbf{w}^\star) &= \lambda\,\text{sign}(\mathbf{w}_j^\star) \;\text{ if }\; \mathbf{w}_j^\star \neq 0, \\ |\mathbf{x}^{j\top}(\mathbf{y}-\mathbf{X}\mathbf{w}^\star)| &\leq \lambda \;\text{ otherwise.} \end{aligned}$$

Uniqueness of the solution

Define $J \triangleq \{j \in \{1,\ldots,p\} : |\mathbf{x}^{j\top}(\mathbf{y}-\mathbf{X}\mathbf{w}^\star)| = \lambda\}$.
If the matrix $\mathbf{X}_J^\top \mathbf{X}_J$ is invertible, the solution is unique and

$$\mathbf{w}_J^\star = (\mathbf{X}_J^\top \mathbf{X}_J)^{-1}(\mathbf{X}_J^\top \mathbf{y} - \lambda \boldsymbol{\eta}_J) = \mathbf{A} + \lambda \mathbf{B},$$

where $\boldsymbol{\eta} \triangleq \text{sign}(\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\mathbf{w}^\star))$.

Brief Introduction to the Homotopy Algorithm

Piecewise linearity

Under uniqueness assumptions of the Lasso solution, the regularization path $\lambda \mapsto \mathbf{w}^\star(\lambda)$ is continuous and piecewise linear.

Brief Introduction to the Homotopy Algorithm

Piecewise linearity

Under uniqueness assumptions of the Lasso solution, the regularization path $\lambda \mapsto \mathbf{w}^\star(\lambda)$ is continuous and piecewise linear.

Recipe of the homotopy method - main ideas

1. finds a trivial solution $\mathbf{w}^\star(\lambda_\infty) = 0$ with $\lambda_\infty = \|\mathbf{X}^\top \mathbf{y}\|_\infty$;
2. compute the direction of the piecewise linear segment of the path;
3. follow the direction of the path by decreasing λ;
4. stop at the next "kink" and go back to 2.

Brief Introduction to the Homotopy Algorithm

Piecewise linearity

Under uniqueness assumptions of the Lasso solution, the regularization path $\lambda \mapsto \mathbf{w}^\star(\lambda)$ is continuous and piecewise linear.

Recipe of the homotopy method - main ideas

1. finds a trivial solution $\mathbf{w}^\star(\lambda_\infty) = 0$ with $\lambda_\infty = \|\mathbf{X}^\top \mathbf{y}\|_\infty$;
2. compute the direction of the piecewise linear segment of the path;
3. follow the direction of the path by decreasing λ;
4. stop at the next "kink" and go back to 2.

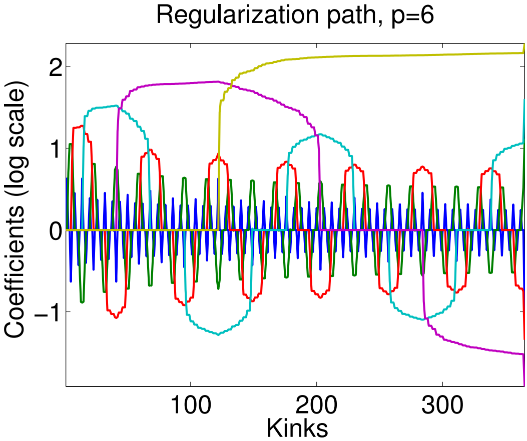
Caveats - questions

- kinks can be very close to each other;
- $\mathbf{X}_J^\top \mathbf{X}_J$ can be ill-conditioned;
- what is the complexity?

Worst case analysis

Theorem - worst case analysis

In the worst-case, the regularization path of the Lasso has exactly $(3^p + 1)/2$ *linear segments.*



Worst case analysis

Consider a Lasso problem ($\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$).
Define the vector $\tilde{\mathbf{y}}$ in $\mathbb{R}^{n+1}$ and the matrix $\tilde{\mathbf{X}}$ in $\mathbb{R}^{(n+1)\times(p+1)}$ as follows:

$$\tilde{\mathbf{y}} \triangleq \begin{bmatrix} \mathbf{y} \\ y_{n+1} \end{bmatrix}, \quad \tilde{\mathbf{X}} \triangleq \begin{bmatrix} \mathbf{X} & 2\alpha\mathbf{y} \\ 0 & \alpha y_{n+1} \end{bmatrix},$$

where $y_{n+1} \neq 0$ and $0 < \alpha < \lambda_1/(2\mathbf{y}^\top\mathbf{y} + y_{n+1}^2)$.

Adverserial strategy

If the regularization path of the Lasso ($\mathbf{y}$,$\mathbf{X}$) has k linear segments, the path of $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}})$ has $3k - 1$ linear segments.

Worst case analysis

$$\tilde{\mathbf{y}} \triangleq \begin{bmatrix} \mathbf{y} \\ y_{n+1} \end{bmatrix}, \quad \tilde{\mathbf{X}} \triangleq \begin{bmatrix} \mathbf{X} & 2\alpha\mathbf{y} \\ 0 & \alpha y_{n+1} \end{bmatrix},$$

Let us denote by $\{\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^k\}$ the sequence of k sparsity patterns in $\{-1, 0, 1\}^p$ encountered along the path of the Lasso $(\mathbf{y}, \mathbf{X})$.

The new sequence of sparsity patterns for $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}})$ is

$$\Bigg\{ \overbrace{\begin{bmatrix} \boldsymbol{\eta}^1 = 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\eta}^2 \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} \boldsymbol{\eta}^k \\ 0 \end{bmatrix}}^{\text{first } k \text{ patterns}}, \overbrace{\begin{bmatrix} \boldsymbol{\eta}^k \\ 1 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\eta}^{k-1} \\ 1 \end{bmatrix}, \ldots, \begin{bmatrix} \boldsymbol{\eta}^1 = 0 \\ 1 \end{bmatrix}}^{\text{middle } k \text{ patterns}},$$

$$\underbrace{\begin{bmatrix} -\boldsymbol{\eta}^2 \\ 1 \end{bmatrix}, \begin{bmatrix} -\boldsymbol{\eta}^3 \\ 1 \end{bmatrix}, \ldots, \begin{bmatrix} -\boldsymbol{\eta}^k \\ 1 \end{bmatrix}}_{\text{last } k\text{–}1 \text{ patterns}} \Bigg\}.$$

Worst case analysis

$$\tilde{\mathbf{y}} \triangleq \begin{bmatrix} \mathbf{y} \\ y_{n+1} \end{bmatrix}, \quad \tilde{\mathbf{X}} \triangleq \begin{bmatrix} \mathbf{X} & 2\alpha\mathbf{y} \\ 0 & \alpha y_{n+1} \end{bmatrix},$$

Some intuition why this is true:

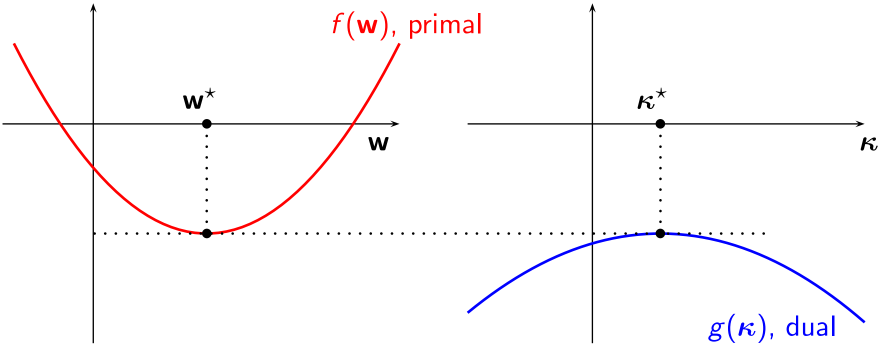
1. the patterns of the new path must be $[\boldsymbol{\eta}^{i\top}, 0]^\top$ or $[\pm\boldsymbol{\eta}^{i\top}, 1]^\top$;
2. the factor α ensures the $(p+1)$-th variable to enter late the path;
3. after the k first kinks, we have $\mathbf{y} \approx \mathbf{X}\mathbf{w}^\star(\lambda)$ and thus

$$\tilde{\mathbf{X}} \begin{bmatrix} \mathbf{w}^\star(\lambda) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ y_{n+1} \end{bmatrix} \approx \tilde{\mathbf{y}} \approx \tilde{\mathbf{X}} \begin{bmatrix} -\mathbf{w}^\star(\lambda) \\ 1/\alpha \end{bmatrix}.$$

We are now in shape to build a pathological path with $(3^p+1)/2$ linear segments. Note that this lower-bound complexity is optimal.

Approximate Complexity

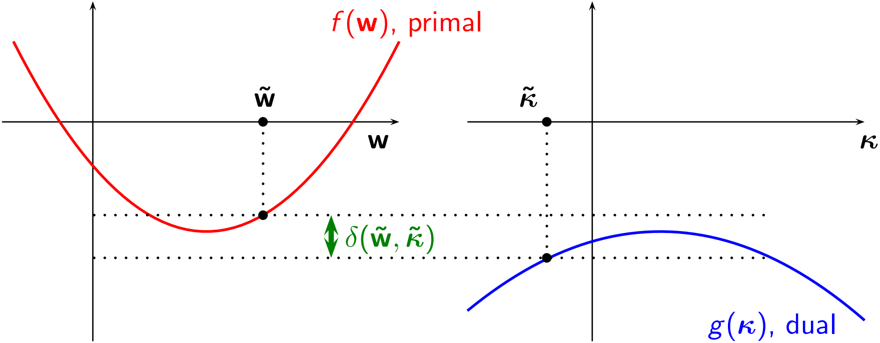
Strong Duality



Strong duality means that $\max_{\boldsymbol{\kappa}} g(\boldsymbol{\kappa}) = \min_{\mathbf{w}} f(\mathbf{w})$

Approximate Complexity

Duality Gaps



$f(\mathbf{w})$, primal
$\tilde{\mathbf{w}}$
$\mathbf{w}$
$\tilde{\boldsymbol{\kappa}}$
$\boldsymbol{\kappa}$
$\delta(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\kappa}})$
$g(\boldsymbol{\kappa})$, dual

Strong duality means that $\max_{\boldsymbol{\kappa}} g(\boldsymbol{\kappa}) = \min_{\mathbf{w}} f(\mathbf{w})$

The duality gap guarantees us that $0 \leq f(\tilde{\mathbf{w}}) - f(\mathbf{w}^\star) \leq \delta(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\kappa}})$.

Approximate Complexity

$$\min_{\mathbf{w}} \left\{ f_\lambda(\mathbf{w}) \triangleq \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1 \right\}, \qquad \text{(primal)}$$

$$\max_{\boldsymbol{\kappa}} \left\{ g_\lambda(\boldsymbol{\kappa}) \triangleq -\frac{1}{2}\boldsymbol{\kappa}^\top\boldsymbol{\kappa} - \boldsymbol{\kappa}^\top\mathbf{y} \ \text{ s.t. } \ \|\mathbf{X}^\top\boldsymbol{\kappa}\|_\infty \leq \lambda \right\}. \qquad \text{(dual)}$$

ε-approximate solution

$\mathbf{w}$ is a ε-approximate solution when there exists a dual variable $\boldsymbol{\kappa}$ s.t.

$$\delta_\lambda(\mathbf{w}, \boldsymbol{\kappa}) = f_\lambda(\mathbf{w}) - g_\lambda(\boldsymbol{\kappa}) \leq \varepsilon f_\lambda(\mathbf{w}).$$

ε-approximate path

A path $\mathcal{P} : \lambda \mapsto \mathbf{w}(\lambda)$ is an approximate path if it always contains ε-approximate solutions.

(see Giesen et al. [2010] for generic results)

Approximate Complexity

ε-approximate solution

$\mathbf{w}$ satisfies $APPROX_\lambda(\varepsilon)$ when there exists a dual variable $\boldsymbol{\kappa}$ s.t.

$$\delta_\lambda(\mathbf{w}, \boldsymbol{\kappa}) = f_\lambda(\mathbf{w}) - g_\lambda(\boldsymbol{\kappa}) \leq \varepsilon f_\lambda(\mathbf{w}).$$

ε-approximate path

A path $\mathcal{P} : \lambda \mapsto \mathbf{w}(\lambda)$ is an approximate path if it always contains ε-approximate solutions.

(see Giesen et al. [2010] for generic results)

Approximate Complexity

Optimality conditions

$\mathbf{w}$ in $\mathbb{R}^p$ is a solution of (1) if and only if for all j in $\{1, \ldots, p\}$,

$$\begin{aligned} \mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\mathbf{w}) &= \lambda \,\mathrm{sign}(\mathbf{w}_j) \;\text{ if }\; \mathbf{w}_j \neq 0, \\ |\mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\mathbf{w})| &\leq \lambda \;\text{ otherwise.} \end{aligned} \qquad \text{(exact)}$$

Approximate Complexity

$(\varepsilon_1, \varepsilon_2)$-approximate optimality conditions

w in $\mathbb{R}^p$ satisfies $OPT_\lambda(\varepsilon_1, \varepsilon_2)$ if and only if for all j in $\{1, \dots, p\}$,

$$\begin{aligned} \lambda(1-\varepsilon_2) \leq \mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\mathbf{w})\,\text{sign}(\mathbf{w}_j) \leq \lambda(1+\varepsilon_1) & \text{ if } \mathbf{w}_j \neq 0, \\ |\mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\mathbf{w})| \leq \lambda(1+\varepsilon_1) & \text{ otherwise.} \end{aligned}$$

Relations between OPT_λ and $APPROX_\lambda$

$$\begin{aligned} APPROX_\lambda(0) & \Longrightarrow OPT_\lambda(0,0) \\ & \Longrightarrow OPT_{\lambda(1-\sqrt(\varepsilon)}(\sqrt(\varepsilon)/(1-\sqrt{\varepsilon}), -\sqrt(\varepsilon)/(1-\sqrt{\varepsilon})) \\ & \Longrightarrow APPROX_{\lambda(1-\sqrt{\varepsilon})}(\varepsilon) \end{aligned}$$

Proposition - approximate analysis

there exists an ε-approximate path with at most $\left\lceil \frac{\log(\lambda_\infty/\lambda_1)}{\sqrt{\varepsilon}} \right\rceil$ segments.

Approximate Homotopy

Recipe - main ideas/features

- Maintain $OPT_\lambda(\varepsilon/2, \varepsilon/2)$ instead of $OPT_\lambda(0, 0)$;
- Make steps in λ greater than or equal to $\lambda(1 - \theta\sqrt{\varepsilon})$;
- When the kinks are too close to each other, make a large step and use a first-order method instead;
- Between λ_∞ and λ_1, the number of iterations is upper-bounded by $\left\lceil \frac{\log(\lambda_\infty/\lambda_1)}{\theta\sqrt{\varepsilon}} \right\rceil$.

Conclusion

A few messages

- **Despite an exponential complexity, the homotopy algorithms remains extremely powerful in practice**;
- the main issue of the homotopy algorithm might be its numerical stability;
- when one does not care about precision, the worst-case complexity of the path can significantly reduce.

Advertisement SPAMS toolbox (open-source)

- C++ interfaced with **Matlab, R, Python**.
- proximal gradient methods for ℓ_0, ℓ_1, **elastic-net, fused-Lasso, group-Lasso, tree group-Lasso, tree-ℓ_0, sparse group Lasso, overlapping group Lasso...**
- ...for **square, logistic, multi-class logistic** loss functions.
- handles sparse matrices, provides duality gaps.
- fast implementations of **OMP** and **LARS - homotopy**.
- dictionary learning and matrix factorization (NMF, sparse PCA).
- coordinate descent, block coordinate descent algorithms.
- fast projections onto some convex sets.

Try it! `http://www.di.ens.fr/willow/SPAMS/`

References I

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, volume 1, pages 267–281, 1973.

S. Bakin. *Adaptive regression and model selection in data mining problems*. PhD thesis, 1999.

R.G. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4): 1982–2001, 2010.

D.P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999. 2nd edition.

J.M. Borwein and A.S. Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer, 2006.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

References II

E. Candès and D. L. Donoho. Recovering edges in ill-posed inverse problems: Optimality of curvelet frames. *Annals of Statistics*, 30(3): 784–842, 2002.

E.J. Candès, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted L1 minimization. *Journal of Fourier Analysis and Applications*, 14(5): 877–905, 2008.

V. Cehver, M. Duarte, C. Hedge, and R.G. Baraniuk. Sparse signal recovery using markov random fields. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1999.

J. F. Claerbout and F. Muir. Robust modeling with erratic data. *Geophysics*, 38(5):826–844, 1973.

References III

J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1160–1169, 1985.

M. Do and M. Vertterli. *Contourlets, Beyond Wavelets*. Academic Press, 2003.

D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

M. A. Efroymson. Multiple regression analysis. *Mathematical methods for digital computers*, 9(1):191–203, 1960.

M. Fazel, H. Hindi, and S. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference*, volume 6, pages 4734–4739, 2001.

References IV

I. E Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

G. M. Furnival and R. W. Wilson. Regressions by leaps and bounds. *Technometrics*, 16(4):499–511, 1974.

B. Gärtner, M. Jaggi, and C. Maria. An exponential lower bound on the complexity of regularization paths. *preprint arXiv:0903.4817v2*, 2010.

J. Giesen, M. Jaggi, and S. Laue. Approximating parameterized convex optimization problems. In *Algorithms - ESA*, Lectures Notes Comp. Sci. 2010.

Y. Grandvalet and S. Canu. Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. In *Advances in Neural Information Processing Systems (NIPS)*, 1999.

R. R. Hocking. A Biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics*, 32:1–49, 1976.

References V

J. Huang, Z. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

L. Jacob, G. Obozinski, and J.-P. Vert. Group Lasso with overlaps and graph Lasso. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

R. Jenatton, J-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011.

V. Klee and G. J. Minty. How good is the simplex algorithm? In O. Shisha, editor, *Inequalities*, volume III, pages 159–175. Academic Press, New York, 1972.

E. Le Pennec and S. Mallat. Sparse geometric image representations with bandelets. *IEEE Transactions on Image Processing*, 14(4): 423–438, 2005.

References VI

S. Mallat. *A wavelet tour of signal processing*. Academic press, 2008. 3rd edition.

C. L. Mallows. Choosing variables in a linear regression: A graphical aid. unpublished paper presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas, 1964.

C. L. Mallows. Choosing a subset regression. unpublished paper presented at the Joint Statistical Meeting, Los Angeles, California, 1966.

C. A. Micchelli, J. M. Morales, and M. Pontil. Regularizers for structured sparsity. *Advances in Computational Mathematics*, 38(3): 455–489, 2013.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24:227–234, 1995.

References VII

M. R. Osborne, B. Presnell, and B. A. Turlach. On the Lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–37, 2000.

F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J.P. Vert. Classification of microarray data using gene networks. *BMC bioinformatics*, 8(1):35, 2007.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5): 465–471, 1978.

K. Ritter. Ein verfahren zur lösung parameterabhängiger, nichtlinearer maximum-probleme. *Mathematical Methods of Operations Research*, 6(4):149–166, 1962.

L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4): 259–268, 1992.

References VIII

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

J. M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41(12): 3445–3462, 1993.

E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multiscale transforms. *IEEE Transactions on Information Theory*, 38(2):587–607, 1992.

N. Srebro, J.D.M. Rennie, and T.S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.

H. L. Taylor, S. C. Banks, and J. F. McCoy. Deconvolution with the ℓ_1 norm. *Geophysics*, 44(1):39–52, 1979.

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

References IX

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society Series B*, 67(1):91–108, 2005.

B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005.

D. Wrinch and H. Jeffreys. XLII. on certain fundamental principles of scientific inquiry. *Philosophical Magazine Series 6*, 42(249):369–390, 1921.

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

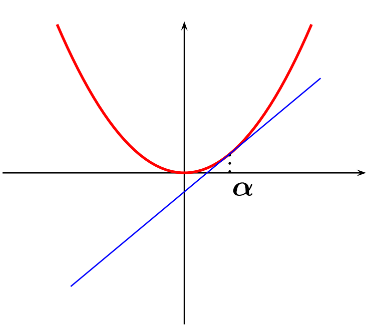
P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37 (6A):3468–3497, 2009.

References X

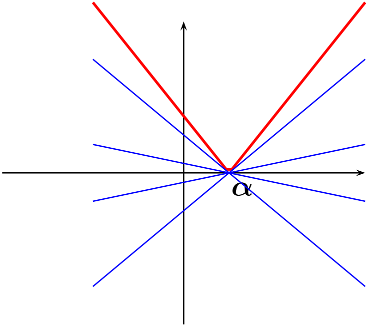
H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2): 301–320, 2005.

Appendix

Basic convex optimization tools: subgradients



(g) Smooth case



(h) Non-smooth case

Figure : Gradients and subgradients for smooth and non-smooth functions.

$$\partial f(\boldsymbol{\alpha}) \stackrel{\triangle}{=} \{\boldsymbol{\kappa} \in \mathbb{R}^p \mid f(\boldsymbol{\alpha}) + \boldsymbol{\kappa}^\top(\boldsymbol{\alpha}' - \boldsymbol{\alpha}) \leq f(\boldsymbol{\alpha}') \text{ for all } \boldsymbol{\alpha}' \in \mathbb{R}^p\}.$$

Basic convex optimization tools: subgradients

Some nice properties

- $\partial f(\boldsymbol{\alpha}) = \{g\}$ iff f differentiable at $\boldsymbol{\alpha}$ and $g = \nabla f(\boldsymbol{\alpha})$.
- many calculus rules: $\partial(\gamma f + \mu g) = \gamma \partial f + \mu \partial g$ for $\gamma, \mu > 0$.

for more details, see Boyd and Vandenberghe [2004], Bertsekas [1999], Borwein and Lewis [2006] and S. Boyd's course at Stanford.

Optimality conditions

For $g : \mathbb{R}^p \to \mathbb{R}$ convex,

- g differentiable: $\boldsymbol{\alpha}^\star$ minimizes g iff $\nabla g(\boldsymbol{\alpha}^\star) = 0$.
- g nondifferentiable: $\boldsymbol{\alpha}^\star$ minimizes g iff $0 \in \partial g(\boldsymbol{\alpha}^\star)$.

Careful: the concept of subgradient requires a function to be above its tangents. It does only make sense for convex functions!

Basic convex optimization tools: dual-norm

Definition

Let $\boldsymbol{\kappa}$ be in $\mathbb{R}^p$,

$$\|\boldsymbol{\kappa}\|_* \triangleq \max_{\boldsymbol{\alpha}\in\mathbb{R}^p : \|\boldsymbol{\alpha}\|\leq 1} \boldsymbol{\alpha}^\top \boldsymbol{\kappa}.$$

Exercises

- $\|\boldsymbol{\alpha}\|_{**} = \|\boldsymbol{\alpha}\|$ (true in finite dimension)
- ℓ_2 is dual to itself.
- ℓ_1 and ℓ_∞ are dual to each other.
- ℓ_q and ℓ_q' are dual to each other if $\frac{1}{q} + \frac{1}{q'} = 1$.
- similar relations for spectral norms on matrices.
- $\partial\|\boldsymbol{\alpha}\| = \{\boldsymbol{\kappa} \in \mathbb{R}^p \;\text{ s.t. }\; \|\boldsymbol{\kappa}\|_* \leq 1 \text{ and } \boldsymbol{\kappa}^\top\boldsymbol{\alpha} = \|\boldsymbol{\alpha}\|\}$.

Optimality conditions

Let $f : \mathbb{R}^p \to \mathbb{R}$ be convex differentiable and $\|.\|$ be any norm.

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} f(\boldsymbol{\alpha}) + \lambda \|\boldsymbol{\alpha}\|.$$

$\boldsymbol{\alpha}$ is solution if and only if

$$0 \in \partial(f(\boldsymbol{\alpha}) + \lambda\|\boldsymbol{\alpha}\|) = \nabla f(\boldsymbol{\alpha}) + \lambda \partial \|\boldsymbol{\alpha}\|$$

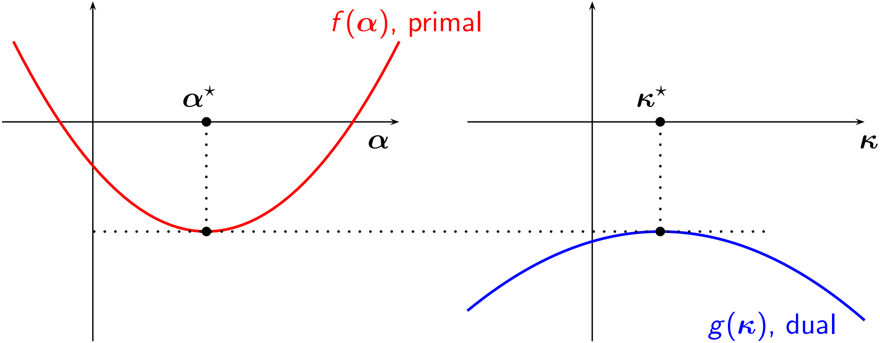
Since $\partial\|\boldsymbol{\alpha}\| = \{\boldsymbol{\kappa} \in \mathbb{R}^p \;\text{ s.t. }\; \|\boldsymbol{\kappa}\|_* \leq 1 \text{ and } \boldsymbol{\kappa}^\top \boldsymbol{\alpha} = \|\boldsymbol{\alpha}\|\}$,

General optimality conditions:

$$\|\nabla f(\boldsymbol{\alpha})\|_* \leq \lambda \;\text{ and }\; -\nabla f(\boldsymbol{\alpha})^\top \boldsymbol{\alpha} = \lambda \|\boldsymbol{\alpha}\|.$$

Convex Duality

Strong Duality

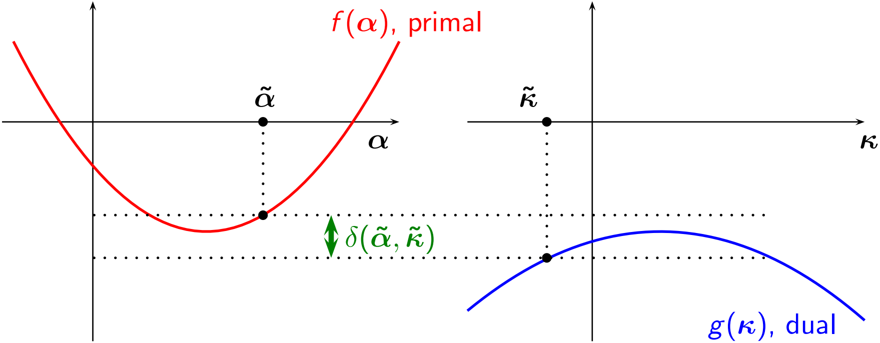


$f(\boldsymbol{\alpha})$, primal
$\boldsymbol{\alpha}^\star$
$\boldsymbol{\alpha}$
$\boldsymbol{\kappa}^\star$
$\boldsymbol{\kappa}$
$g(\boldsymbol{\kappa})$, dual

Strong duality means that $\max_{\boldsymbol{\kappa}} g(\boldsymbol{\kappa}) = \min_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha})$

Convex Duality

Duality Gaps


$f(\boldsymbol{\alpha})$, primal
$\tilde{\boldsymbol{\alpha}}$
$\boldsymbol{\alpha}$
$\tilde{\boldsymbol{\kappa}}$
$\boldsymbol{\kappa}$
$\delta(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\kappa}})$
$g(\boldsymbol{\kappa})$, dual

Strong duality means that $\max_{\boldsymbol{\kappa}} g(\boldsymbol{\kappa}) = \min_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha})$

The duality gap guarantees us that $0 \leq f(\tilde{\boldsymbol{\alpha}}) - f(\boldsymbol{\alpha}^\star) \leq \delta(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\kappa}})$.