Optimization for Sparse Estimation and Structured Sparsity

Julien Mairal

INRIA LEAR, Grenoble

IMA, Minneapolis, June 2013

Short Course "Applied Statistics and Machine Learning"

Main topic of the lecture

How do we solve

- sparse estimation problems in machine learning and statistics;
- sparse inverse problems in signal processing.

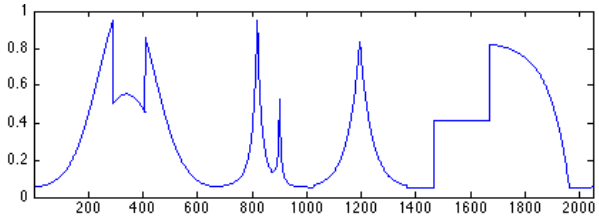
Tools we use

- greedy algorithms;
- "gradient" methods for non-differentiable functions;
- other tricks.

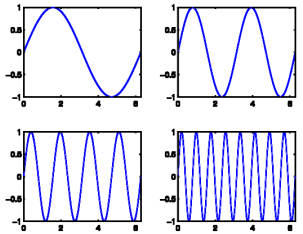
Part I: Short Introduction on Sparsity

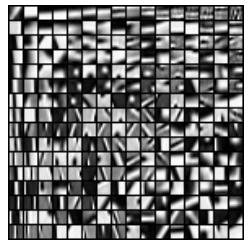
Sparse Linear Models: Signal Processing Point of View

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



Let $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^p] \in \mathbb{R}^{n \times p}$ be a set of elementary signals.
We call it **dictionary**.



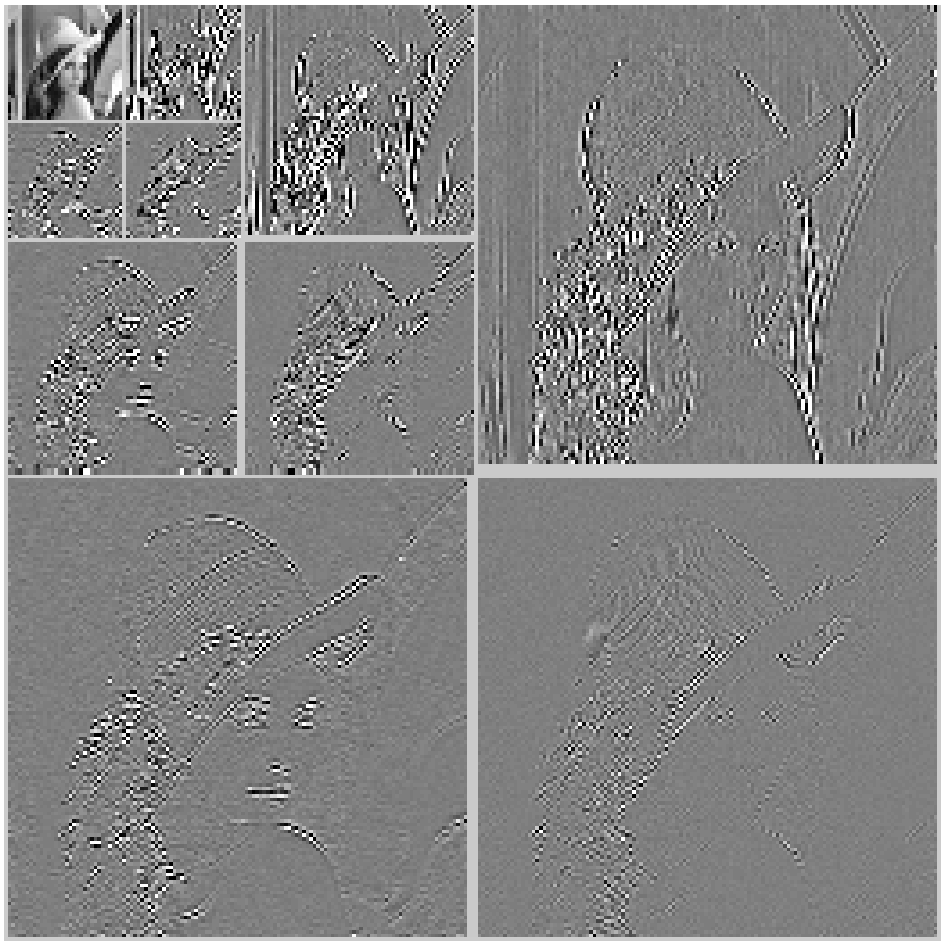


$\mathbf{X}$ is "adapted" to $\mathbf{y}$ if it can represent it with a few elements—that is, there exists a **sparse vector** β in $\mathbb{R}^p$ such that $\mathbf{y} \approx \mathbf{X}\beta$. We call β the **sparse code**.

$$\underbrace{\begin{pmatrix} \mathbf{y} \end{pmatrix}}_{\mathbf{y} \in \mathbb{R}^n} \approx \underbrace{\left(\begin{array}{c|c|c|c} \mathbf{x}^1 & \mathbf{x}^2 & \cdots & \mathbf{x}^p \end{array} \right)}_{\mathbf{X} \in \mathbb{R}^{n \times p}} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}}_{\beta \in \mathbb{R}^p, \textbf{sparse}}$$

Sparse Linear Models: Signal Processing Point of View

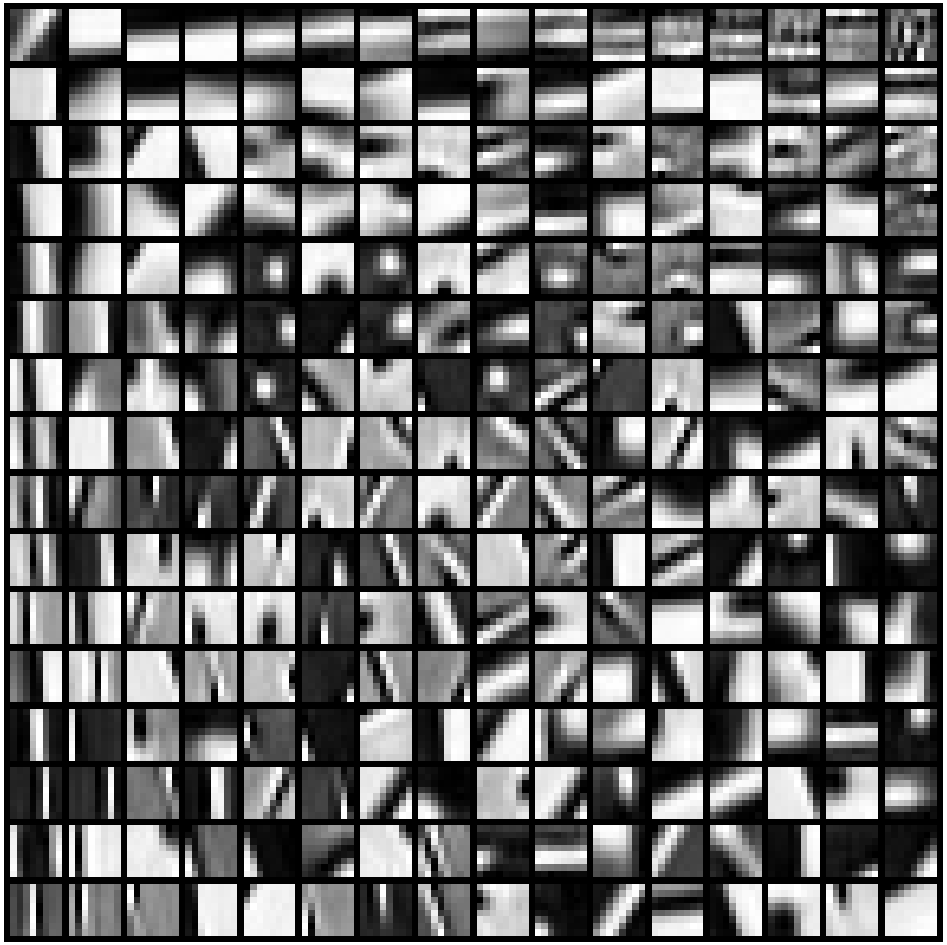
Wavelet Representation [see Mallat, 1999]



$\mathbf{y} = \mathbf{X}\beta$. The coefficients β are represented in a quadtree.

Sparse Linear Models: Signal Processing Point of View

Dictionary for natural image patches [Olshausen and Field, 1997, Elad and Aharon, 2006]



Sparse Linear Models: Machine Learning 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 is a relation $y \approx \beta^\top \mathbf{x}$, and solve

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

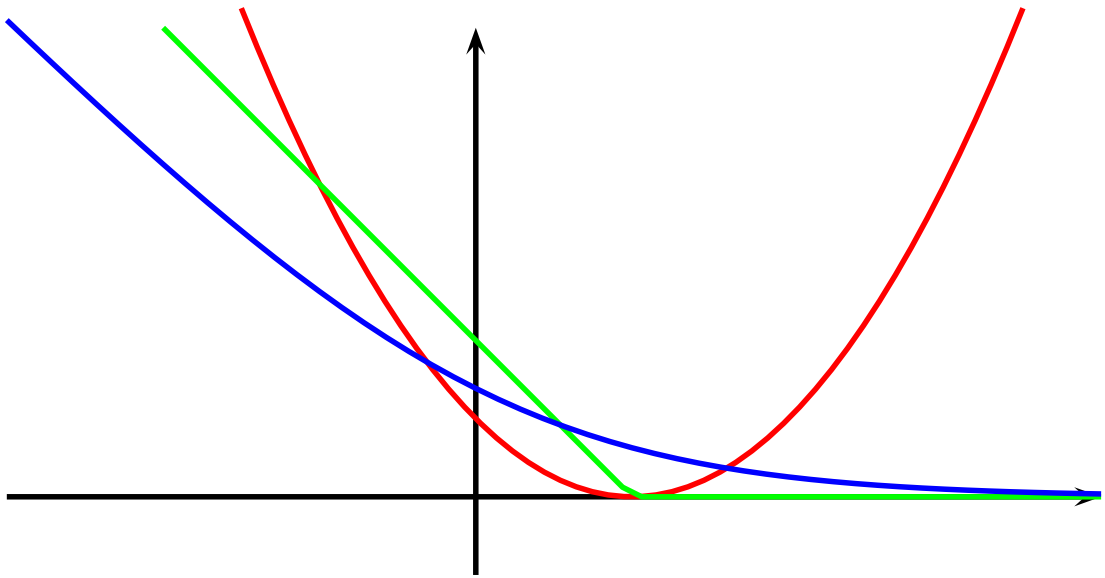
Sparse Linear Models: Machine Learning Point of View

A few examples:

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

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

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



Sparse Linear Models: Machine Learning Point of View

A few examples:

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

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

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

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

Main Topic of the lecture

How do we solve?

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \underbrace{\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}_{\text{loss/data-fitting}} + \underbrace{\lambda \Omega(\boldsymbol{\beta})}_{\text{regularization}},$$

$\mathbf{X}$ in $\mathbb{R}^{n \times p}$ is a design matrix, Ω induces sparsity.

- $\|\boldsymbol{\beta}\|_0 \triangleq \#\{i \text{ s.t. } \boldsymbol{\beta}_i \neq 0\}$ (NP-hard)
- $\|\boldsymbol{\beta}\|_1 \triangleq \sum_{i=1}^p |\boldsymbol{\beta}_i|$ (convex),

When Ω is the ℓ_1-norm, the problem is called Lasso [Tibshirani, 1996] or basis pursuit [Chen et al., 1999].

Later in the lecture, extensions to more general smooth loss functions f, other sparsity-inducing penalties.

One Important Question

Why does the ℓ_1-norm induce sparsity?

Can we get some intuition from the simplest isotropic case?

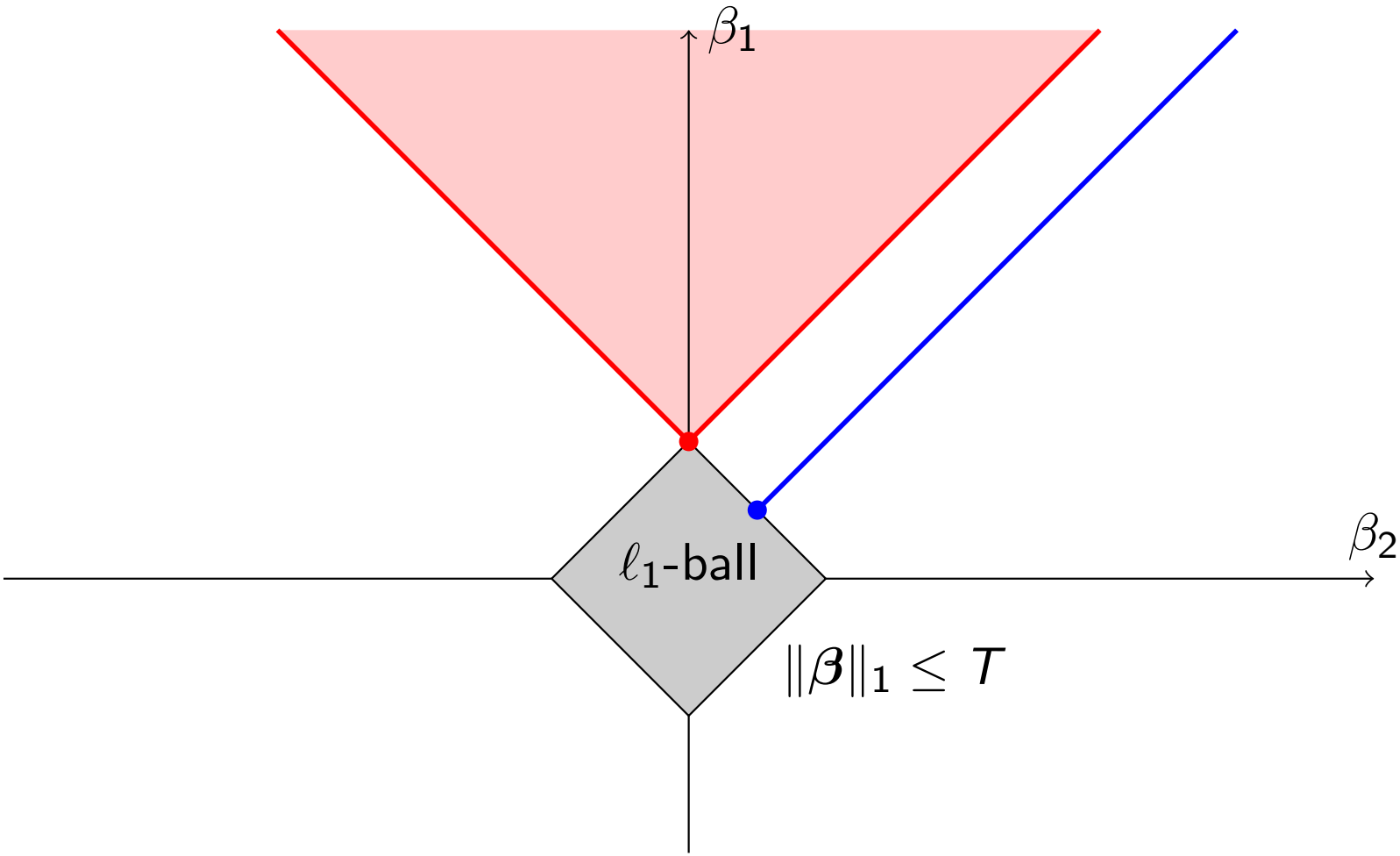
$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y}-\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

or equivalently the Euclidean projection onto the ℓ_1-ball?

$$\tilde{\boldsymbol{\beta}}(T) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y}-\boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \leq T.$$

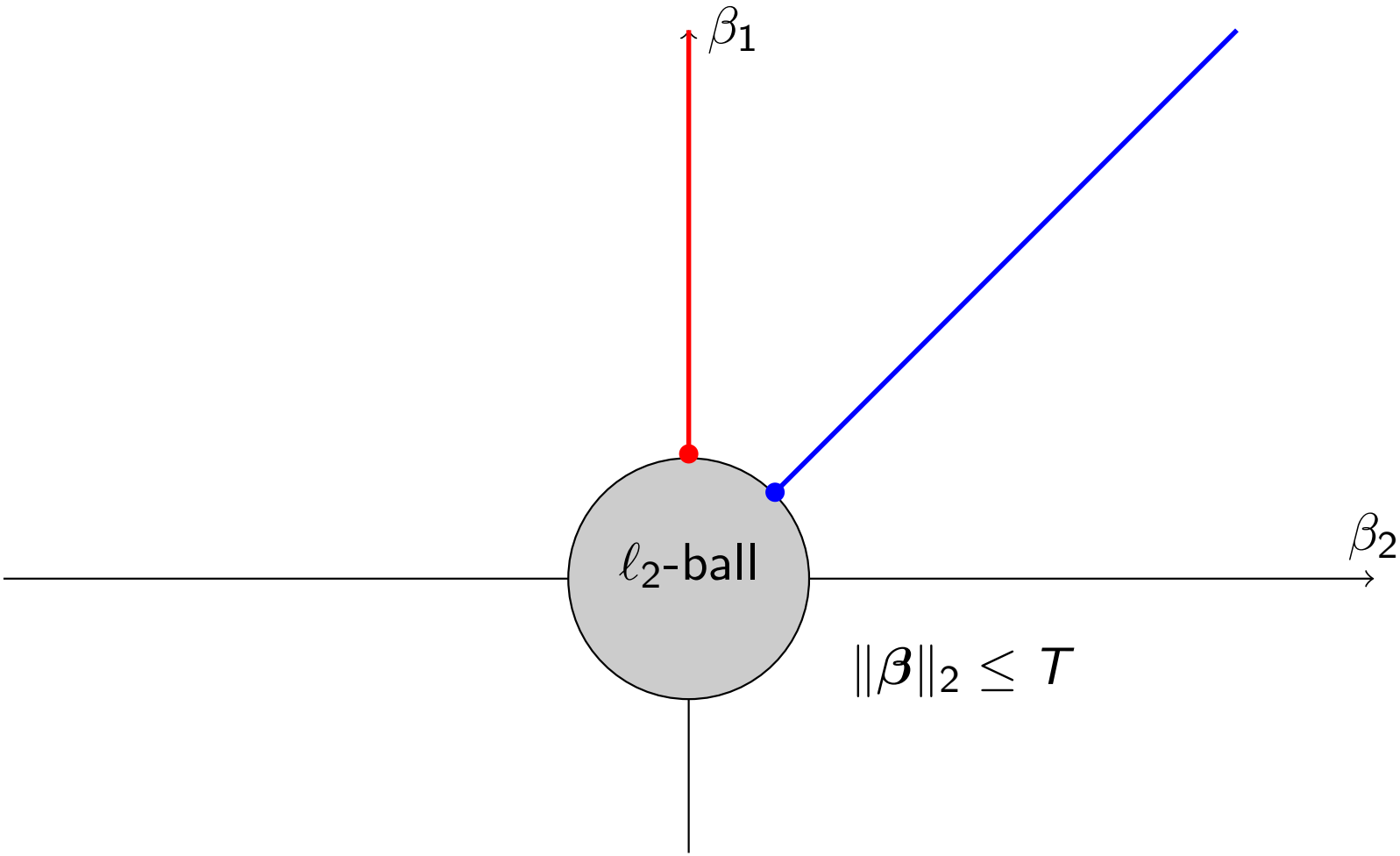
"equivalent" means that $\forall\lambda > 0, \exists T \geq 0$ s.t. $\tilde{\boldsymbol{\beta}}(T) = \hat{\boldsymbol{\beta}}(\lambda)$.

Regularizing with the ℓ_1-norm



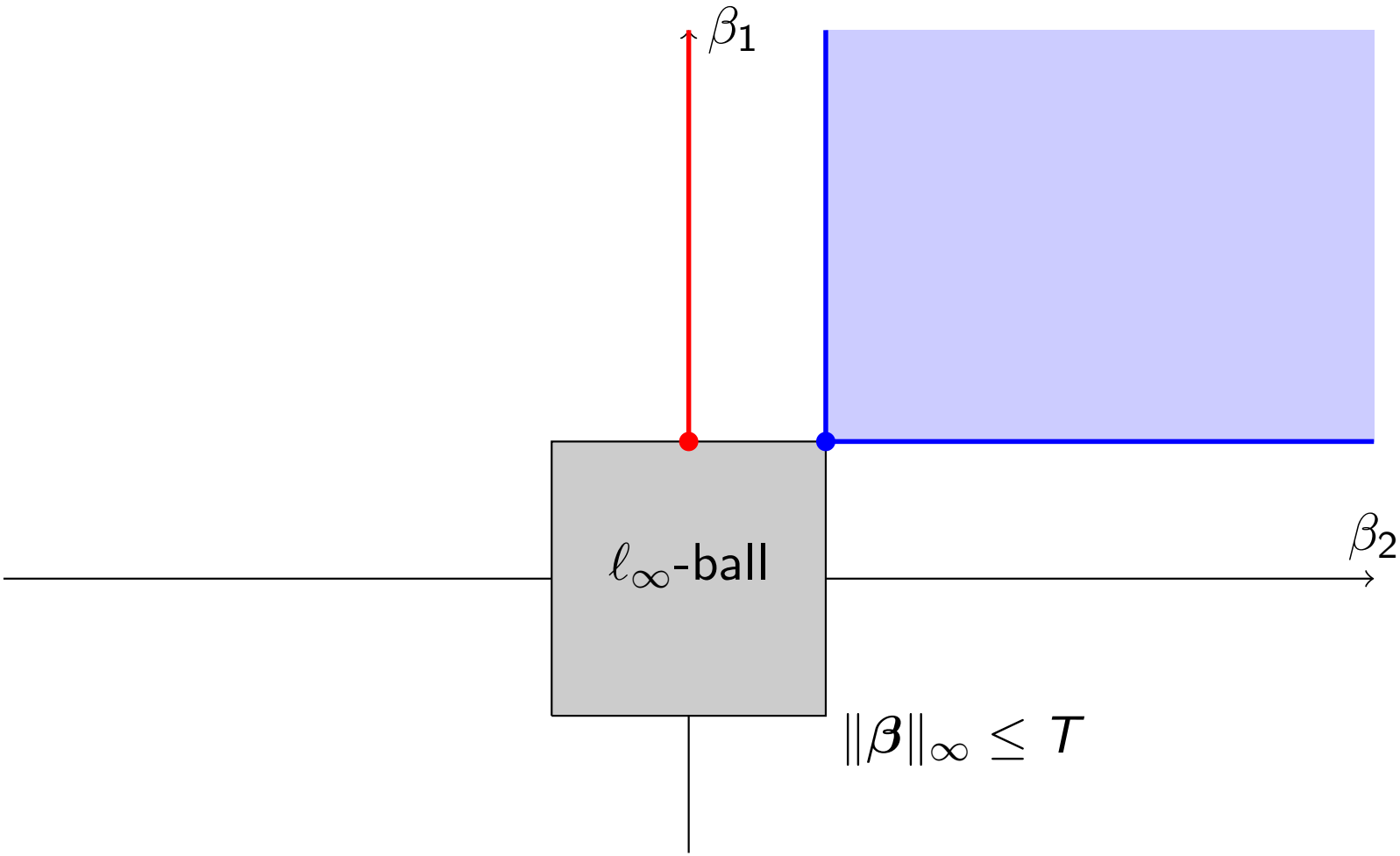
The projection onto a convex set is "biased" towards singularities.

Regularizing with the ℓ_2-norm



The ℓ_2-norm is isotropic.

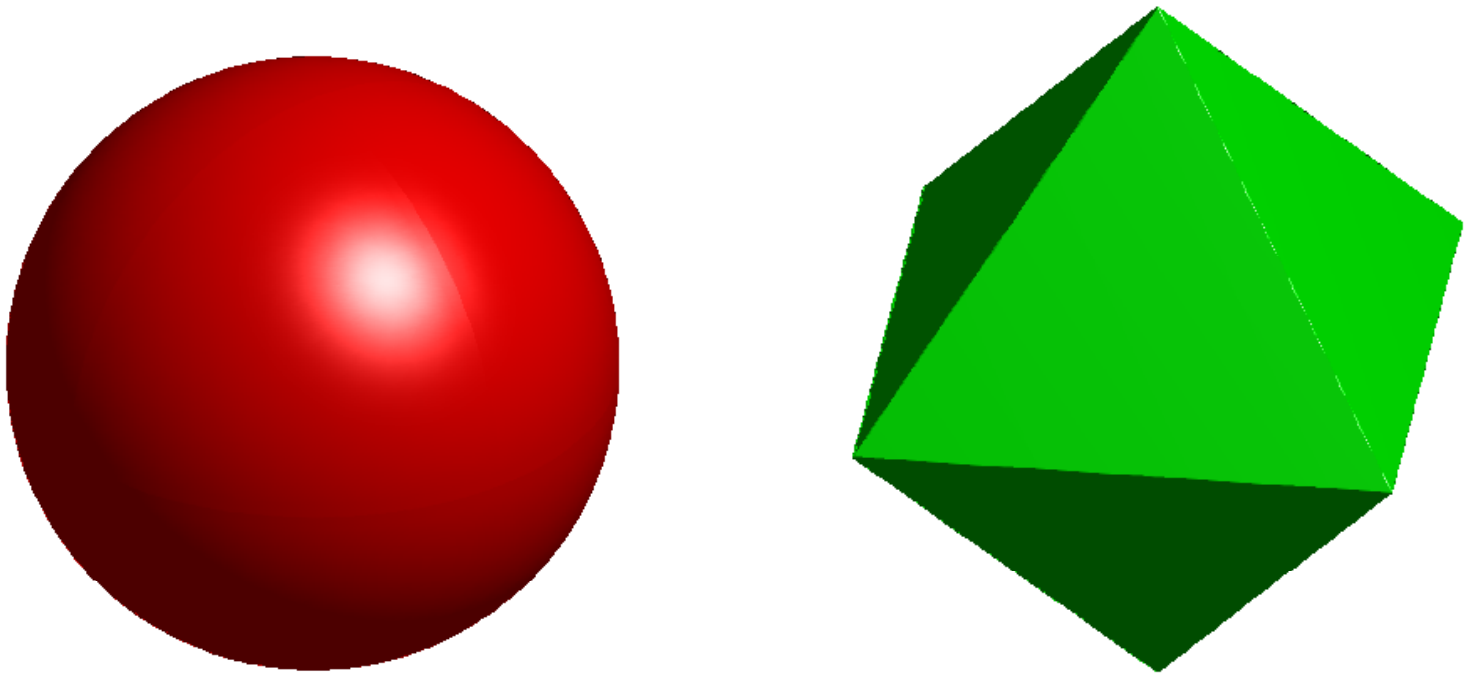
Regularizing with the ℓ_∞-norm



The ℓ_∞-norm encourages $|\beta_1| = |\beta_2|$.

In 3D.

Copyright G. Obozinski



Why does the ℓ_1-norm induce sparsity?

Exemple: quadratic problem in 1D

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

Piecewise quadratic function with a kink at zero.

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

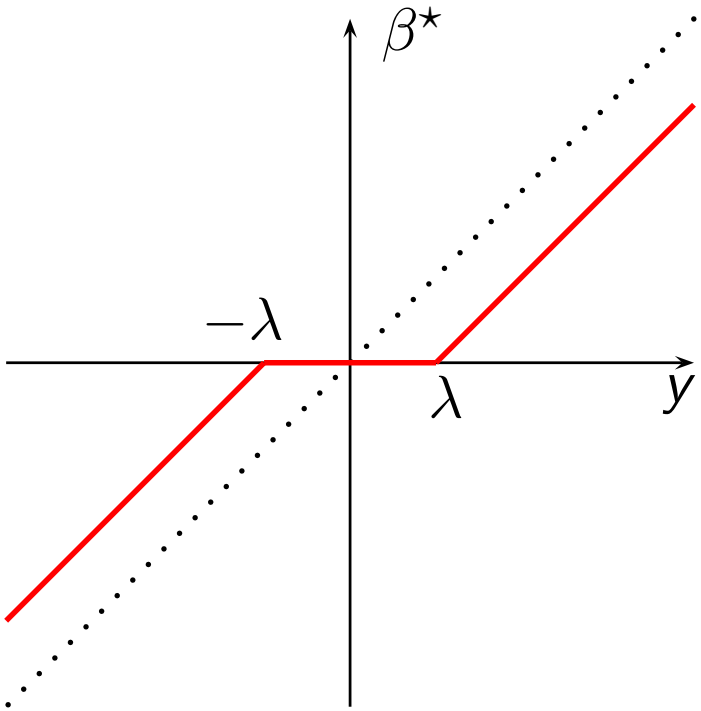
Optimality conditions. β is optimal iff:

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

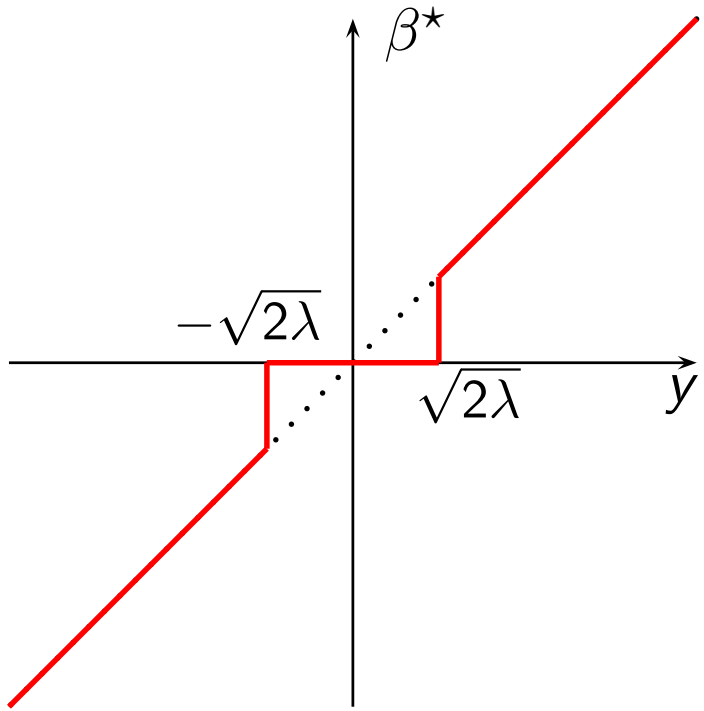
The solution is a **soft-thresholding**:

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

Why does the ℓ_1-norm induce sparsity?



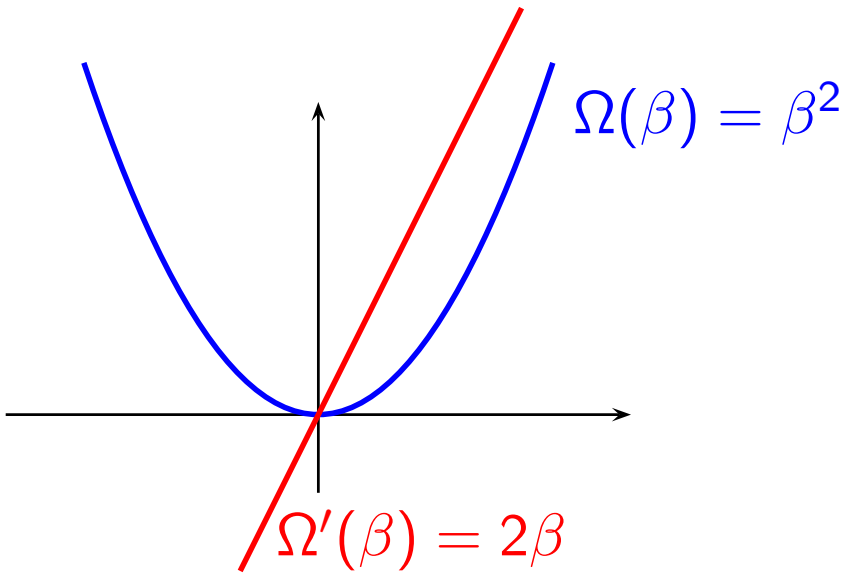
(a) soft-thresholding operator,
$\beta^\star = \text{sign}(y)(|y| - \lambda)^+$,
$\min_\beta \frac{1}{2}(y-\beta)^2 + \lambda|\beta|$

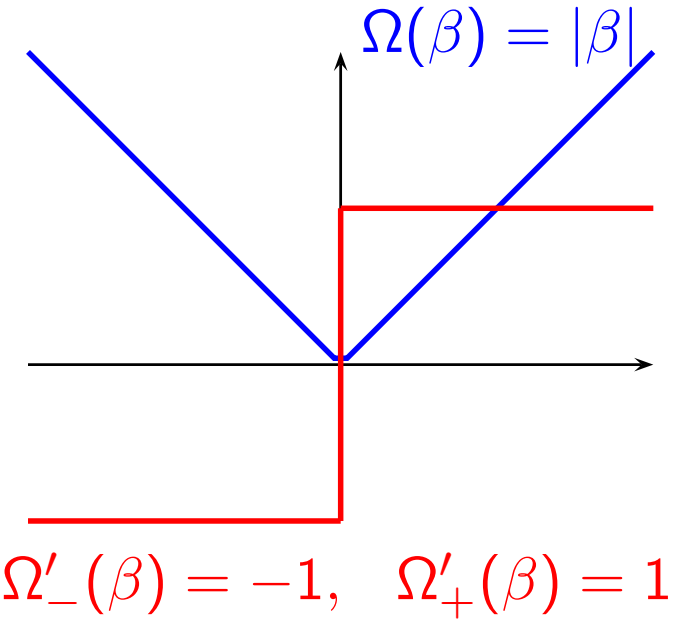


(b) hard-thresholding operator
$\beta^\star = \mathbf{1}_{|u| \geq \sqrt{2\lambda}} y$
$\min_\beta \frac{1}{2}(y-\beta)^2 + \lambda \mathbf{1}_{|\beta|>0}$

Why does the ℓ_1-norm induce sparsity?

Comparison with ℓ_2-regularization in 1D

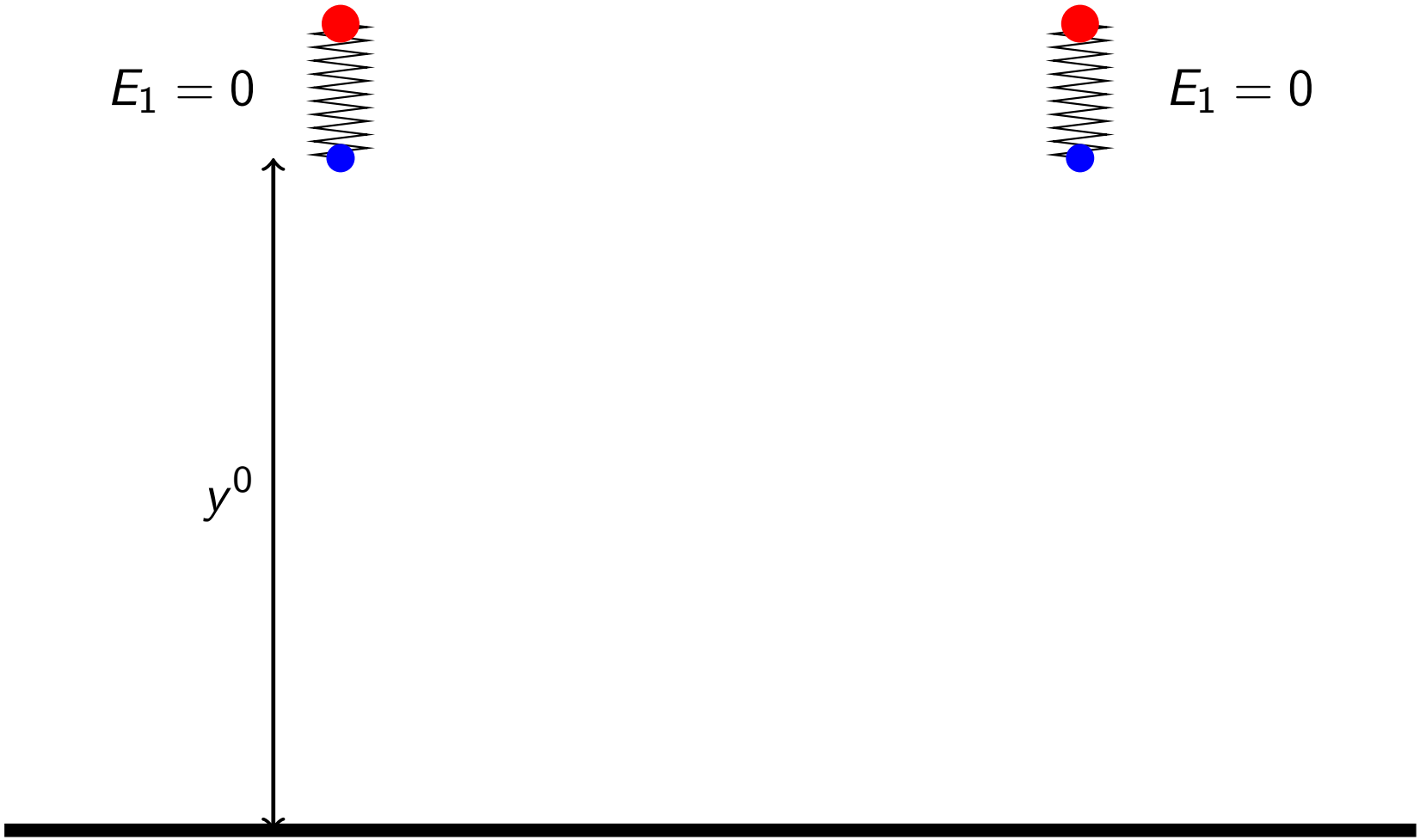




The gradient of the ℓ_2-penalty vanishes when β get close to 0. On its differentiable part, the norm of the gradient of the ℓ_1-norm is constant.

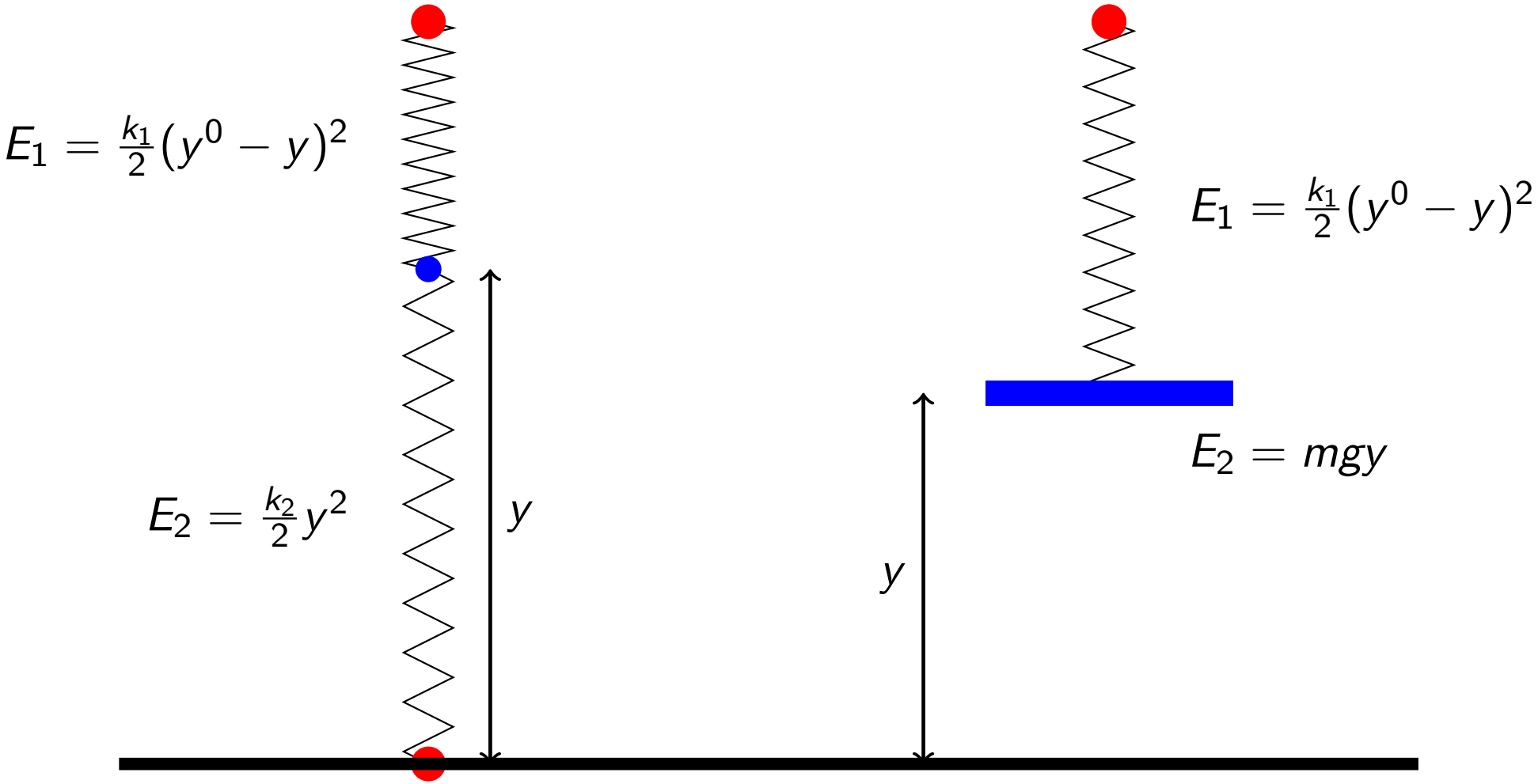
Why does the ℓ_1-norm induce sparsity?

Physical illustration



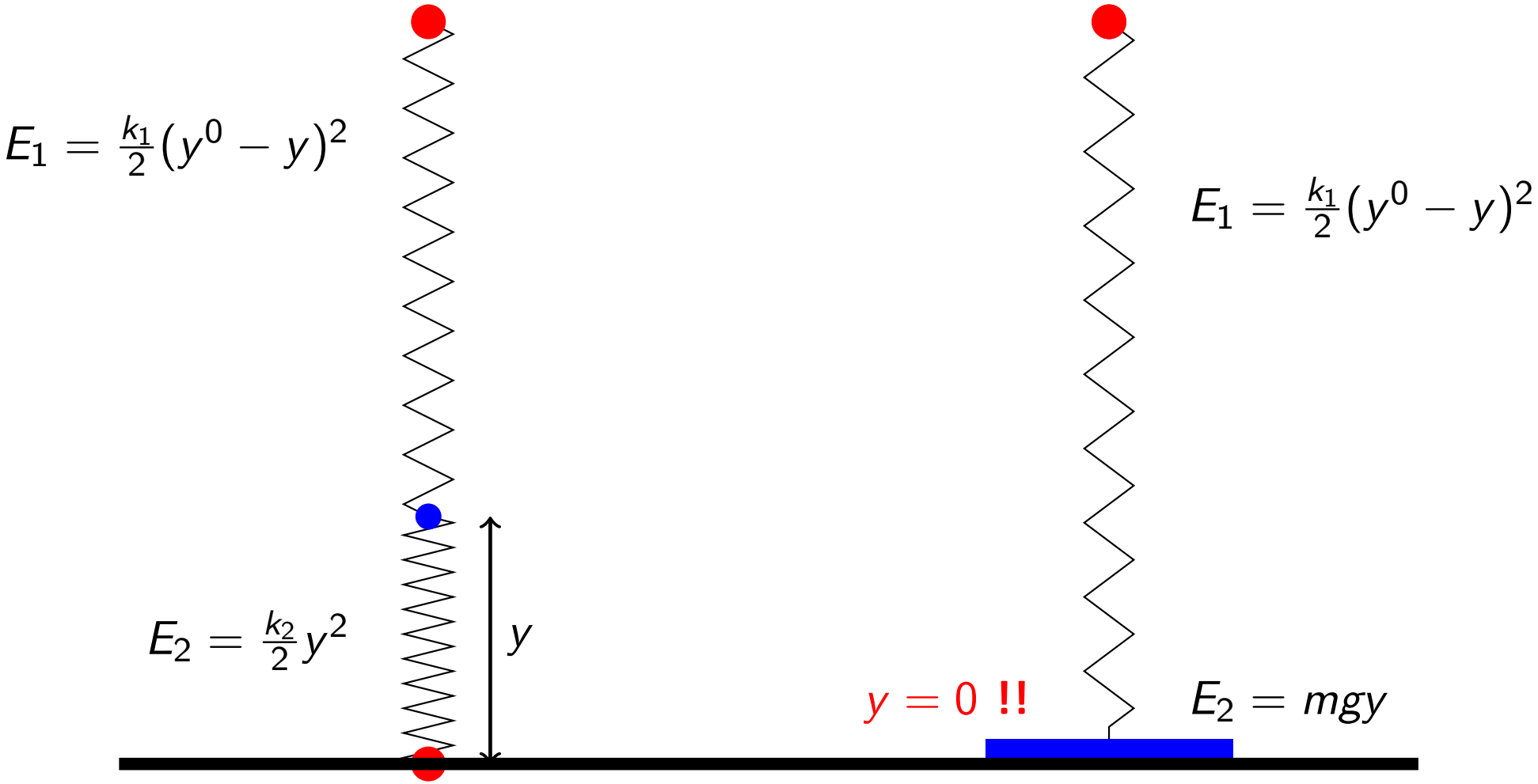
Why does the ℓ_1-norm induce sparsity?

Physical illustration



Why does the ℓ_1-norm induce sparsity?

Physical illustration



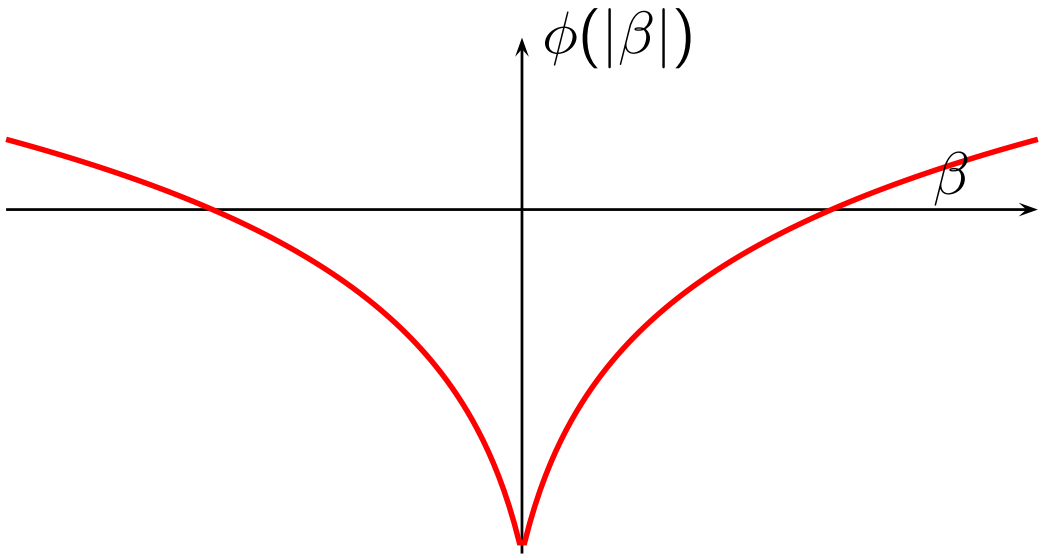
Examples of sparsity-inducing penalties

Exploiting concave functions with a kink at zero

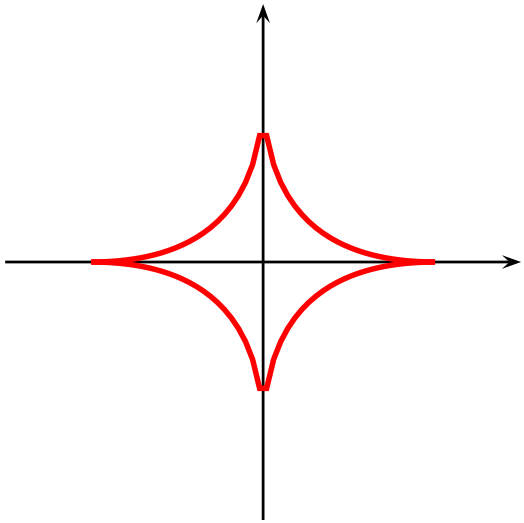
$\Omega(\beta) = \sum_{j=1}^{p} \phi(|\beta_j|)$.

- ℓ_q-"pseudo-norm", with $0 < q < 1$: $\Omega(\beta) \triangleq \sum_{j=1}^{p} |\beta_j|^q$,
- log penalty, $\Omega(\beta) \triangleq \sum_{j=1}^{p} \log(|\beta_j| + \varepsilon)$,

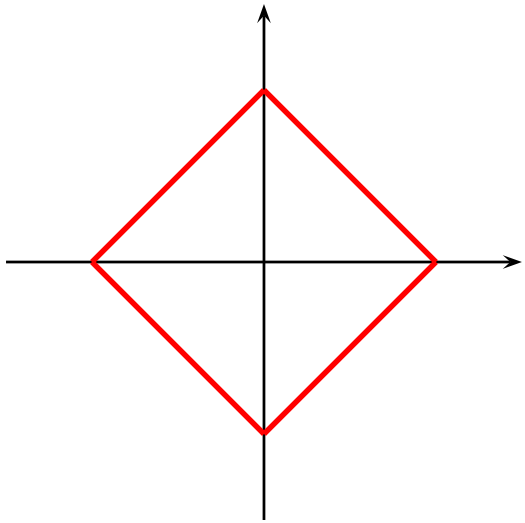
ϕ is any function that looks like this:



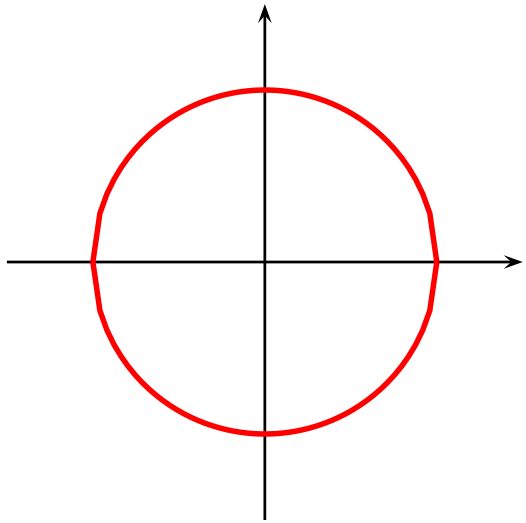
Examples of sparsity-inducing penalties



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



(d) ℓ_1-ball, 2-D



(e) ℓ_2-ball, 2-D

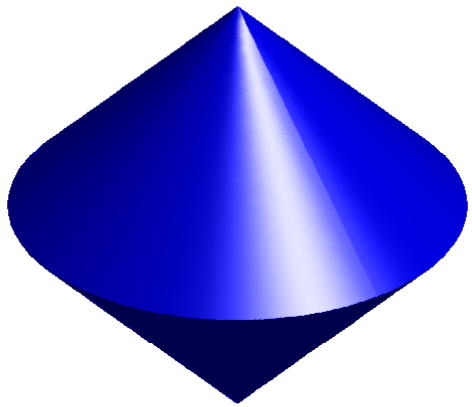
Figure: Open balls in 2-D corresponding to several ℓ_q-norms and pseudo-norms.

Group Lasso

Turlach et al. [2005], Yuan and Lin [2006], Zhao et al. [2009]

$$\text{the } \ell_1/\ell_q\text{-norm}: \qquad \Omega(\beta) = \sum_{g \in \mathcal{G}} \|\beta_g\|_q.$$

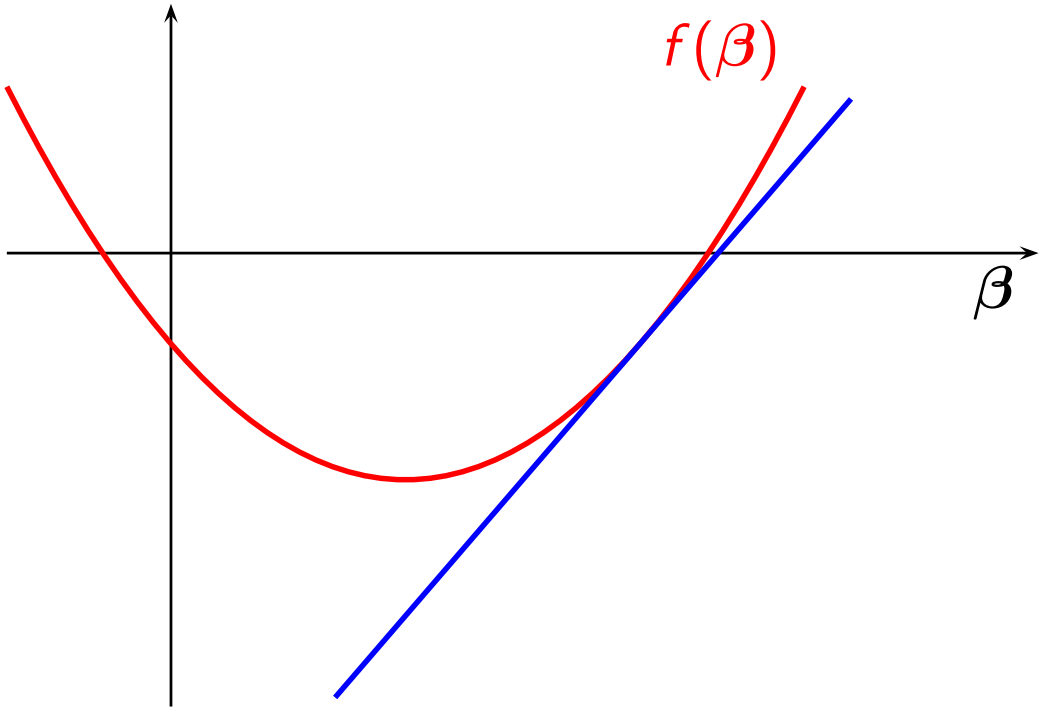
- $\mathcal{G}$ is a **partition** of $\{1, \ldots, p\}$;
- $q = 2$ or $q = \infty$ in practice;
- can be interpreted as the ℓ_1-norm of $[\|\beta_g\|_q]_{g \in \mathcal{G}}$.



$$\Omega(\beta) = \|\beta_{\{1,2\}}\|_2 + |\beta_3|.$$

Part II: Convex Optimization for Sparse Estimation

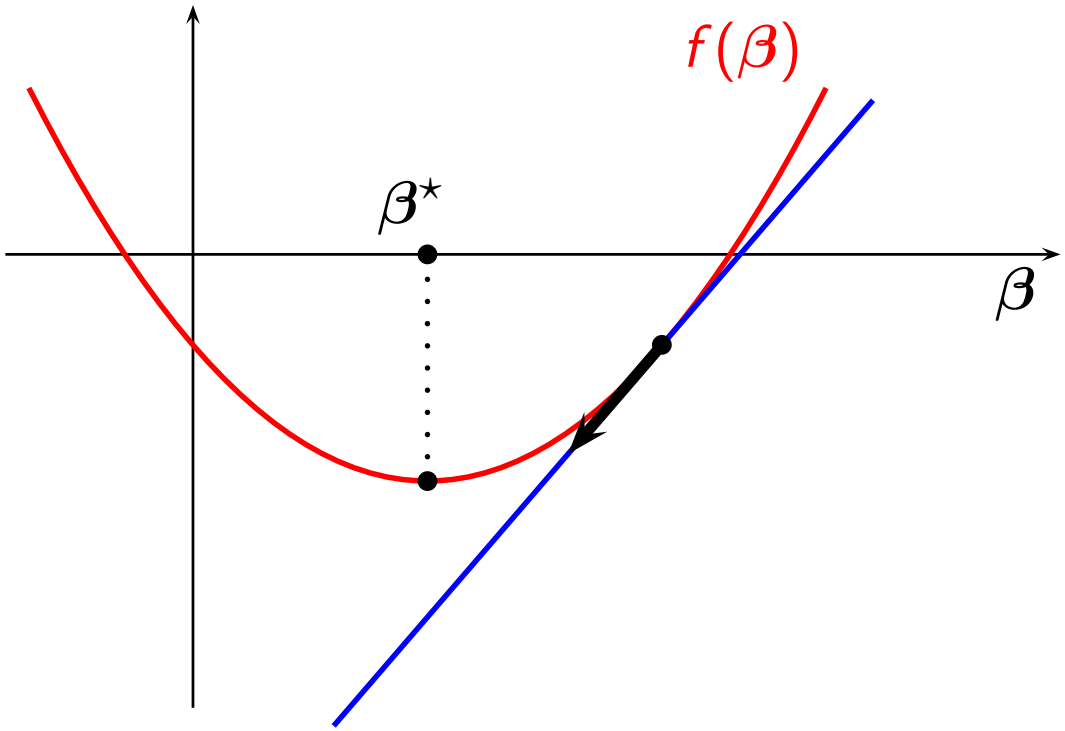
Convex Functions



Why do we care about convexity?

Why do we care about convexity?

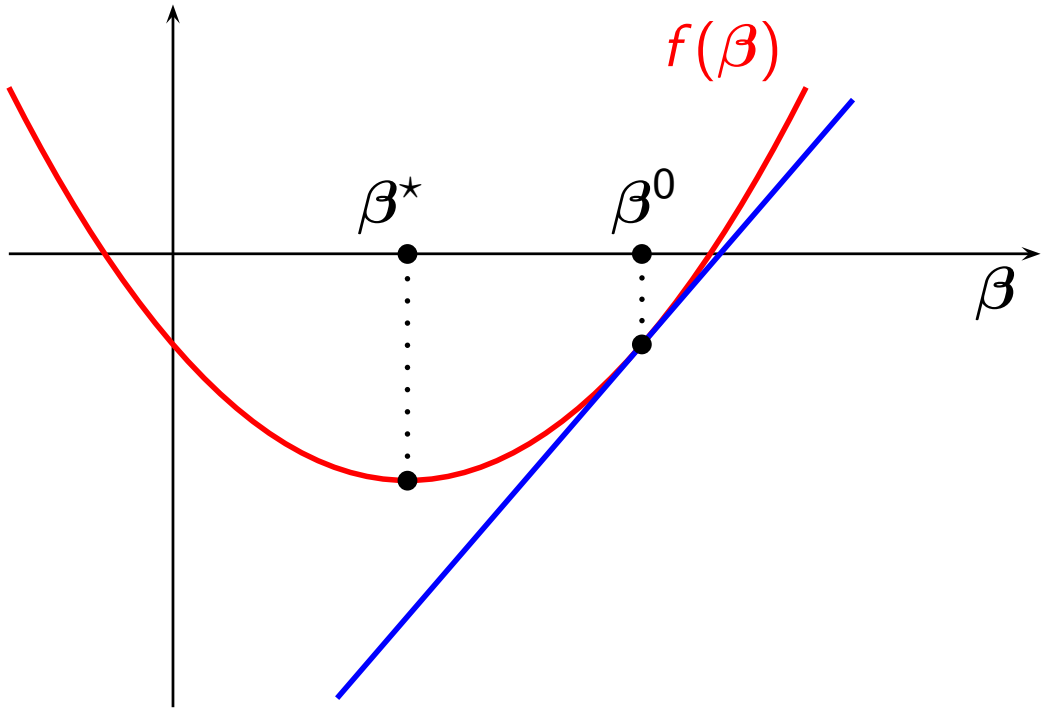
Local observations give some information about the global optimum



- $\nabla f(\beta) = 0$ is a necessary and sufficient optimality condition for differentiable convex functions;
- it is often easy to upper-bound $f(\beta) - f^\star$.

An important inequality for smooth convex functions

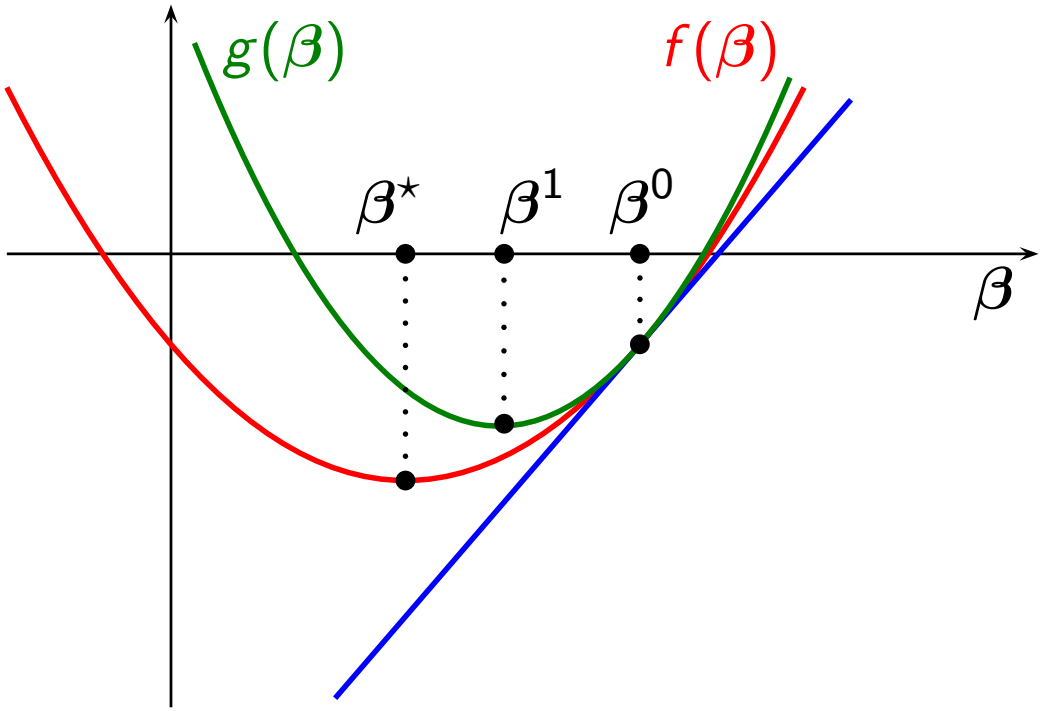
If f is convex



- $f(\beta) \geq \underbrace{f(\beta^0) + \nabla f(\beta^0)^\top (\beta - \beta^0)}_{\text{linear approximation}}$;
- this is an equivalent definition of convexity for smooth functions.

An important inequality for smooth functions

If ∇f is L-Lipschitz continuous



- $f(\beta) \leq g(\beta) = \underbrace{f(\beta^0) + \nabla f(\beta^0)^\top(\beta - \beta^0)}_{\text{linear approximation}} + \frac{L}{2}\|\beta - \beta^0\|_2^2;$
- $\boxed{\beta^1 = \beta^0 - \frac{1}{L}\nabla f(\beta^0). \text{ (gradient descent step).}}$

Gradient Descent Algorithm

Assume that f is convex and differentiable, and that ∇f is L-Lipschitz.

Theorem

Consider the algorithm

$$\boxed{\beta^t \leftarrow \beta^{t-1} - \frac{1}{L}\nabla f(\beta^{t-1}).}$$

Then,

$$f(\beta^t) - f^\star \leq \frac{L\|\beta^0 - \beta^\star\|_2^2}{2t}.$$

Remarks

- the convergence rate improves under additional assumptions on f (strong convexity);
- some variants have a $O(1/t^2)$ convergence rate [Nesterov, 1983].

Proof (1/2)

Proof of the main inequality for smooth functions

We want to show that for all $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$,

$$f(\boldsymbol{\beta}) \leq f(\boldsymbol{\alpha}) + \nabla f(\boldsymbol{\alpha})^\top (\boldsymbol{\beta} - \boldsymbol{\alpha}) + \frac{L}{2}\|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2^2.$$

By using Taylor's theorem with integral form,

$$f(\boldsymbol{\beta}) - f(\boldsymbol{\alpha}) = \int_0^1 \nabla f(t\boldsymbol{\beta} + (1-t)\boldsymbol{\alpha})^\top (\boldsymbol{\beta} - \boldsymbol{\alpha}) dt.$$

Then,

$$\begin{aligned}
f(\boldsymbol{\beta}) - f(\boldsymbol{\alpha}) - \nabla f(\boldsymbol{\alpha})^\top (\boldsymbol{\beta} - \boldsymbol{\alpha}) &\leq \int_0^1 (\nabla f(t\boldsymbol{\beta} + (1-t)\boldsymbol{\alpha}) - \nabla f(\boldsymbol{\alpha}))^\top (\boldsymbol{\beta} - \boldsymbol{\alpha}) dt \\
&\leq \int_0^1 |(\nabla f(t\boldsymbol{\beta} + (1-t)\boldsymbol{\alpha}) - \nabla f(\boldsymbol{\alpha}))^\top (\boldsymbol{\beta} - \boldsymbol{\alpha})| dt \\
&\leq \int_0^1 \|\nabla f(t\boldsymbol{\beta} + (1-t)\boldsymbol{\alpha}) - \nabla f(\boldsymbol{\alpha})\|_2 \|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2 dt \quad \text{(C.-S.)} \\
&\leq \int_0^1 Lt\|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2^2 dt = \frac{L}{2}\|\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2^2.
\end{aligned}$$

Proof (2/2)

Proof of the theorem

We have shown that for all $\boldsymbol{\beta}$,

$$f(\boldsymbol{\beta}) \leq g_t(\boldsymbol{\beta}) = f(\boldsymbol{\beta}^{t-1}) + \nabla f(\boldsymbol{\beta}^{t-1})^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^{t-1}) + \frac{L}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^{t-1}\|_2^2.$$

g_t is minimized by $\boldsymbol{\beta}^t$; it can be rewritten $g_t(\boldsymbol{\beta}) = g_t(\boldsymbol{\beta}^t) + \frac{L}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^t\|_2^2$. Then,

$$\begin{aligned}
f(\boldsymbol{\beta}^t) \leq g_t(\boldsymbol{\beta}^t) &= g_t(\boldsymbol{\beta}^\star) - \frac{L}{2}\|\boldsymbol{\beta}^\star - \boldsymbol{\beta}^t\|_2^2 \\
&= f(\boldsymbol{\beta}^{t-1}) + \nabla f(\boldsymbol{\beta}^{t-1})^\top (\boldsymbol{\beta}^\star - \boldsymbol{\beta}^{t-1}) + \frac{L}{2}\|\boldsymbol{\beta}^\star - \boldsymbol{\beta}^{t-1}\|_2^2 - \frac{L}{2}\|\boldsymbol{\beta}^\star - \boldsymbol{\beta}^t\|_2^2 \\
&\leq f^\star + \frac{L}{2}\|\boldsymbol{\beta}^\star - \boldsymbol{\beta}^{t-1}\|_2^2 - \frac{L}{2}\|\boldsymbol{\beta}^\star - \boldsymbol{\beta}^t\|_2^2.
\end{aligned}$$

By summing from $t = 1$ to T, we have a telescopic sum

$$T(f(\boldsymbol{\beta}^T) - f^\star) \leq \sum_{t=1}^{T} f(\boldsymbol{\beta}^t) - f^\star \leq \frac{L}{2}\|\boldsymbol{\beta}^\star - \boldsymbol{\beta}^0\|_2^2 - \frac{L}{2}\|\boldsymbol{\beta}^\star - \boldsymbol{\beta}^T\|_2^2.$$

The Proximal Gradient Method

We consider a smooth function f and a non-smooth regularizer Ω.

$$\min_{\beta \in \mathbb{R}^p} f(\beta) + \Omega(\beta)$$

For example,

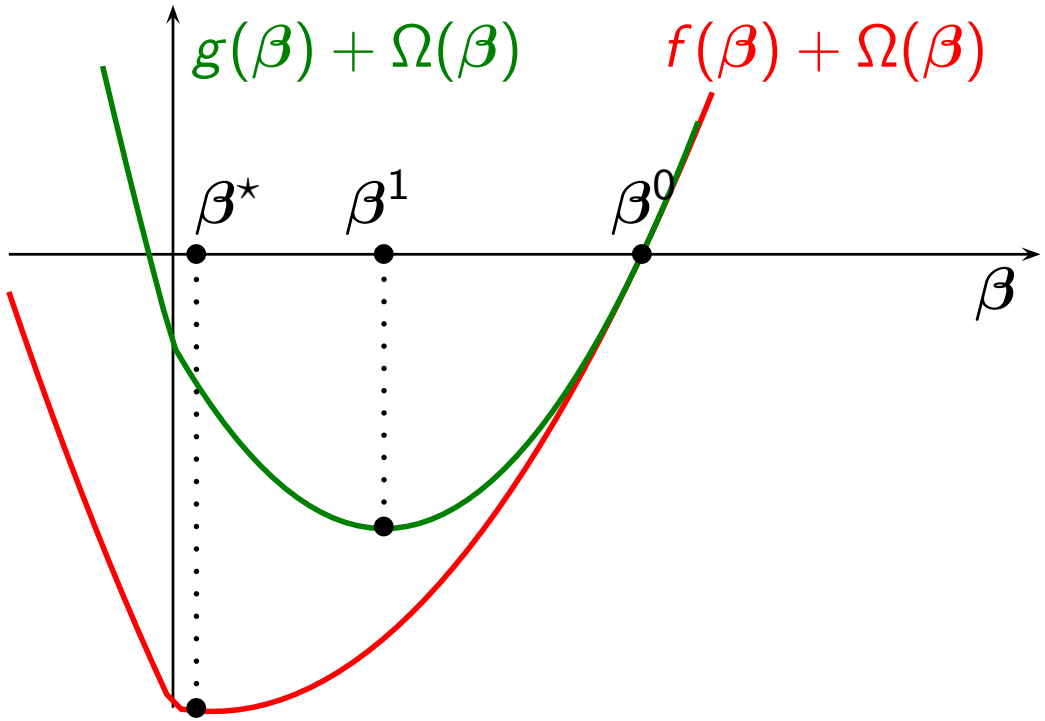
$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1.$$

- the objective function is not differentiable.
- an extension of gradient descent for such a problem is called **"proximal gradient descent"**
  [Beck and Teboulle, 2009, Nesterov, 2007].

The Proximal Gradient Method

An important inequality for composite functions

If ∇f is L-Lipschitz continuous



- $f(\beta) + \Omega(\beta) \leq f(\beta^0) + \nabla f(\beta^0)^\top(\beta - \beta^0) + \frac{L}{2}\|\beta - \beta^0\|_2^2 + \Omega(\beta)$;
- β^1 minimizes $g + \Omega$.

The Proximal Gradient Method

Gradient descent for minimizing f consists of

$$\beta^t \leftarrow \arg\min_{\beta \in \mathbb{R}^p} g_t(\beta) \quad \Longleftrightarrow \quad \beta^t \leftarrow \beta^{t-1} - \frac{1}{L}\nabla f(\beta^{t-1}).$$

The proximal gradient method for minimizing $f + \Omega$ consists of

$$\beta^t \leftarrow \arg\min_{\beta \in \mathbb{R}^p} g_t(\beta) + \Omega(\beta),$$

which is equivalent to

$$\boxed{\beta^t \leftarrow \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\left\|\beta^{t-1} - \frac{1}{L}\nabla f(\beta^{t-1}) - \beta\right\|_2^2 + \frac{1}{L}\Omega(\beta).}$$

It requires computing efficiently the **proximal operator** of Ω.

$$\alpha \mapsto \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|\alpha - \beta\|_2^2 + \Omega(\beta).$$

The Proximal Gradient Method

Remarks

- also known as forward-backward algorithm;
- has similar convergence rates as the gradient descent method [Beck and Teboulle, 2009, Nesterov, 2007];
- there exists line search schemes to automatically tune L.

The case of ℓ_1

The proximal operator of $\lambda\|.\|_1$ is the soft-thresholding operator

$$\boldsymbol{\beta}_j^\star = \text{sign}(\boldsymbol{\alpha}_j)(|\boldsymbol{\alpha}_j| - \lambda)^+.$$

The resulting algorithm is called iterative soft-thresholding [Daubechies et al., 2004].

The Proximal Gradient Method for the Group Lasso

The proximal operator for the Group Lasso penalty

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{\alpha}-\boldsymbol{\beta}\|_2^2 + \lambda \sum_{g\in\mathcal{G}} \|\boldsymbol{\beta}_g\|_q.$$

For $q=2$,

$$\boldsymbol{\beta}_g^\star = \frac{\boldsymbol{\alpha}_g}{\|\boldsymbol{\alpha}_g\|_2}(\|\boldsymbol{\alpha}_g\|_2-\lambda)^+, \quad \forall g\in\mathcal{G}.$$

For $q=\infty$,

$$\boldsymbol{\beta}_g^\star = \boldsymbol{\alpha}_g - \Pi_{\|.\|_1\leq\lambda}[\boldsymbol{\alpha}_g], \quad \forall g\in\mathcal{G}.$$

These formula generalize soft-thresholding to groups of variables.

Coordinate Descent for the Lasso

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1.$$

The coordinate descent method consists of iteratively fixing all variables and optimizing with respect to one:

$$\beta_j \leftarrow \arg\min_{\beta\in\mathbb{R}} \frac{1}{2}\| \underbrace{\mathbf{y}-\sum_{i\neq j}\boldsymbol{\beta}_i\mathbf{x}^i}_{\mathbf{r}} -\beta\mathbf{x}^j\|_2^2+\lambda|\beta|,$$

where $\mathbf{X}=[\mathbf{x}^1,\ldots,\mathbf{x}^p]$. Assume the columns of $\mathbf{X}$ to have unit ℓ_2-norm,

$$\boldsymbol{\beta}_j \leftarrow \text{sign}(\mathbf{x}^{j\top}\mathbf{r})(|\mathbf{x}^{j\top}\mathbf{r}|-\lambda)^+$$

This involves again the **soft-thresholding** operator.

Coordinate Descent for the Lasso

Remarks

- no parameter to tune!
- impressive performance with five lines of code.
- coordinate descent + nonsmooth objective is **not convergent in general**. Here, the problem is equivalent to a convex smooth optimization problem with separable constraints

$$\min_{\beta_+,\beta_-} \frac{1}{2}\|\mathbf{y} - \mathbf{X}_+\beta_+ + \mathbf{X}_-\beta_-\|_2^2 + \lambda\beta_+^T\mathbf{1} + \lambda\beta_-^T\mathbf{1} \quad \text{s.t.} \quad \beta_-, \beta_+ \geq 0.$$

  For this specific problem, the algorithm is **convergent** [see Bertsekas, 1999].
- can be extended to group-Lasso, or other loss functions.
- j can be picked up at random, or by cycling (harder to analyze).

Smoothing Techniques: Reweighted ℓ_2

Let us start from something simple

$$a^2 - 2ab + b^2 \geq 0.$$

Smoothing Techniques: Reweighted ℓ_2

Let us start from something simple

$$a^2 - 2ab + b^2 \geq 0.$$

Then

$$a \leq \frac{1}{2}\left(\frac{a^2}{b} + b\right) \quad \text{with equality iff} \quad a = b$$

and

$$\|\beta\|_1 = \min_{\eta_j \geq 0} \frac{1}{2} \sum_{j=1}^{p} \frac{\beta_j^2}{\eta_j} + \eta_j.$$

The formulation becomes

$$\min_{\beta, \eta_j \geq \varepsilon} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \frac{\lambda}{2} \sum_{j=1}^{p} \frac{\beta_j^2}{\eta_j} + \eta_j.$$

The Regularization Path of the Lasso

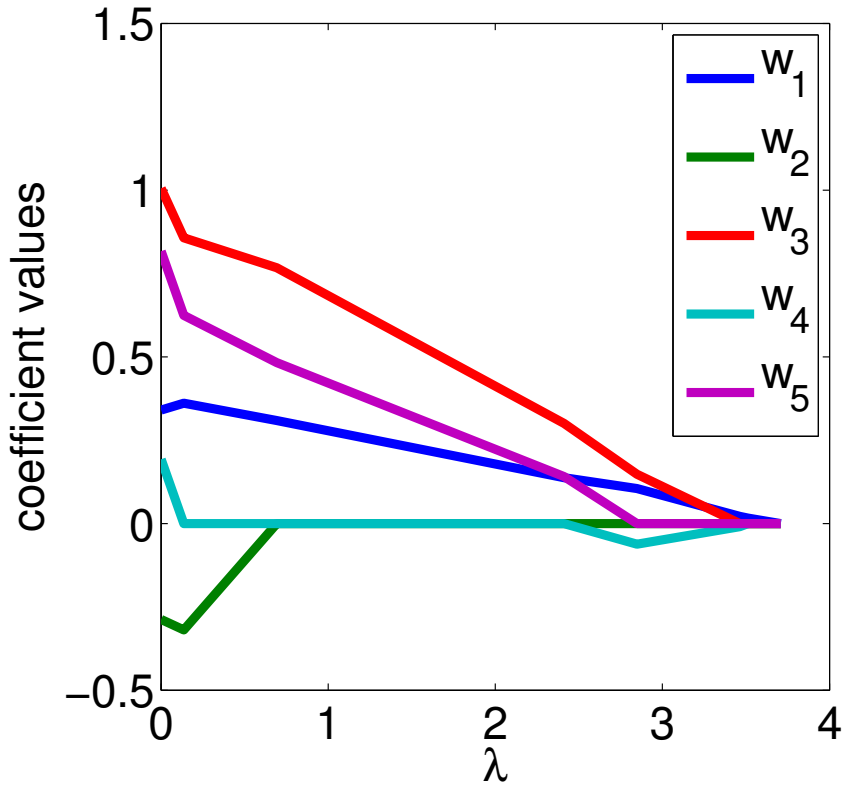


Figure: The regularization path of the Lasso is piecewise linear.

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

Lasso and optimality conditions

Theorem

β is a solution of the Lasso if and only if

$$\left\{ \begin{array}{rcll} |\mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\beta)| & \leq & \lambda & \text{if } \beta_j = 0 \\ \mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\beta) & = & \lambda \operatorname{sign}(\beta_j) & \text{otherwise.} \end{array} \right.$$

Consequence

$$\beta^\star_\Gamma(\lambda) = (\mathbf{X}_\Gamma^T \mathbf{X}_\Gamma)^{-1}(\mathbf{X}_\Gamma^T \mathbf{y} - \lambda \operatorname{sign}(\beta^\star_\Gamma)) = \mathbf{A} + \lambda \mathbf{B},$$

where $\Gamma = \{j \text{ s.t. } \beta_j \neq 0\}$. If we know Γ and the signs of $\beta^\star$ in advance, we have a closed form solution.

Following the piecewise linear regularization path is called the **homotopy** method [Osborne et al., 2000, Efron et al., 2004].

LARS algorithm (Homotopy)

The regularization path $(\lambda, \beta^\star(\lambda))$ is piecewise linear.

1. Start from the trivial solution $(\lambda = \|\mathbf{X}^T\mathbf{y}\|_\infty, \beta^\star(\lambda) = 0)$.
2. Define $\Gamma = \{j \text{ s.t. } |\mathbf{x}^{j\top}\mathbf{y}| = \lambda\}$,
3. Follow the regularization path: $\beta^\star_\Gamma(\lambda) = \mathbf{A} + \lambda\mathbf{B}$, keeping $\beta^\star_{\Gamma^c} = 0$, decreasing the value of λ, until one of the following event occurs:
   - $\exists j \notin \Gamma$ such that $|\mathbf{x}^{j\top}(\mathbf{y} - \mathbf{X}\beta^\star(\lambda))| = \lambda$, then $\Gamma \leftarrow \Gamma \cup \{j\}$.
   - $\exists j \in \Gamma$ such that $\beta^\star(\lambda) = 0$, then $\Gamma \leftarrow \Gamma \setminus \{j\}$.
4. Update the direction of the path and go back to 3.

Hidden assumptions

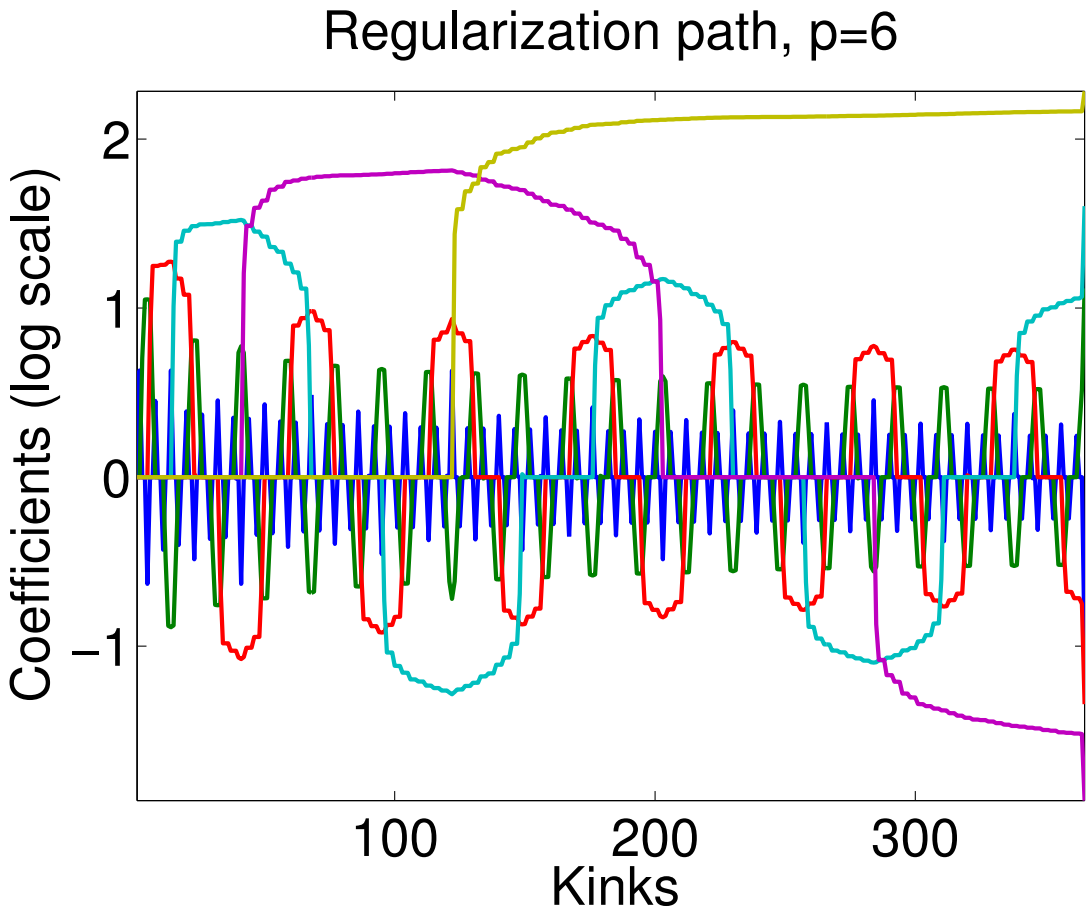
- the regularization path is unique.
- variables enter the path one at a time.

Extremely efficient for **small/medium scale** problems ($p \leq 10\,000$) and/or **very sparse** problems (when implemented correctly). **Robust to correlated features**. Can solve the elastic-net.

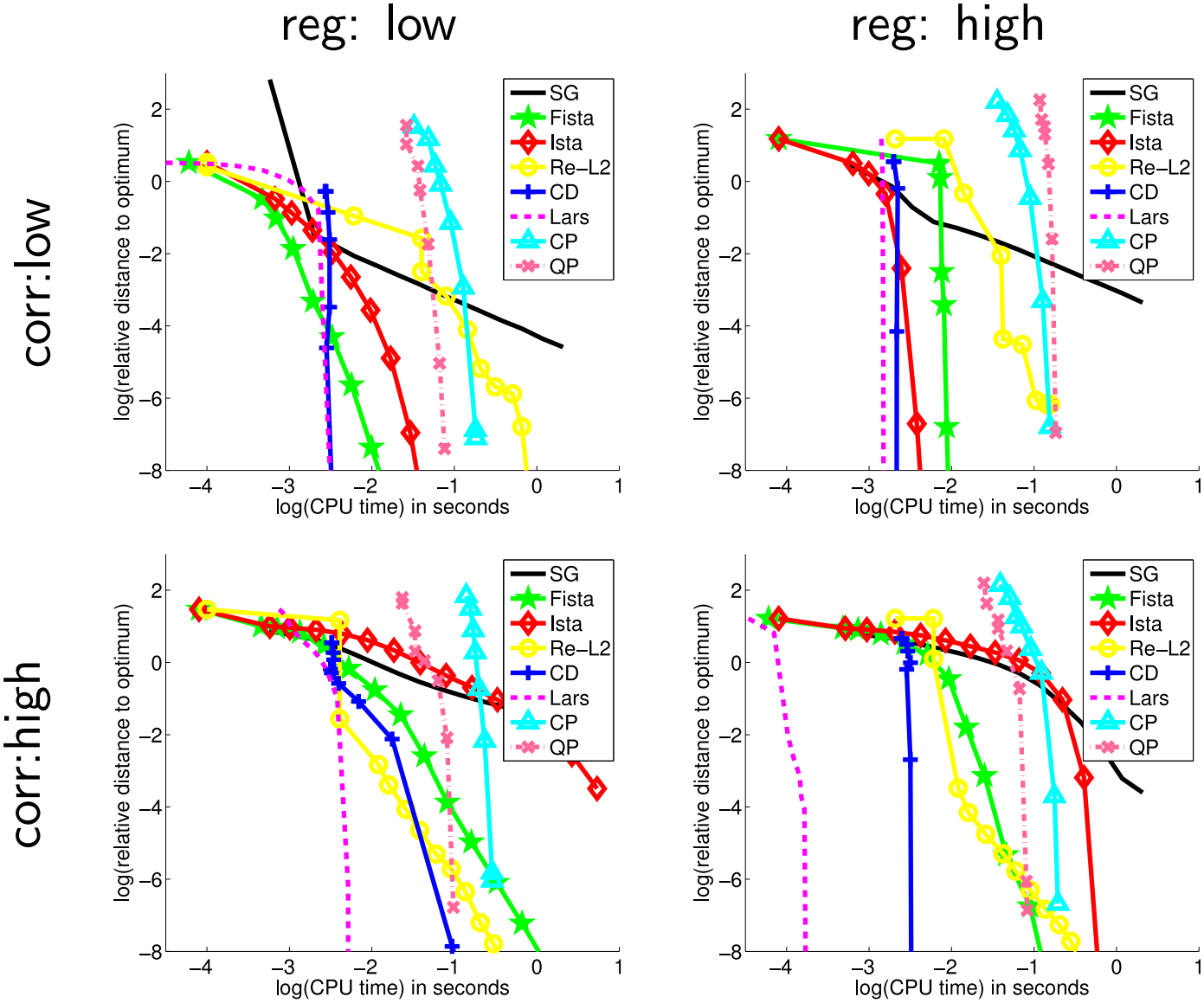
LARS algorithm (Homotopy) - Complexity

Theorem - worst case analysis [Mairal and Yu, 2012]

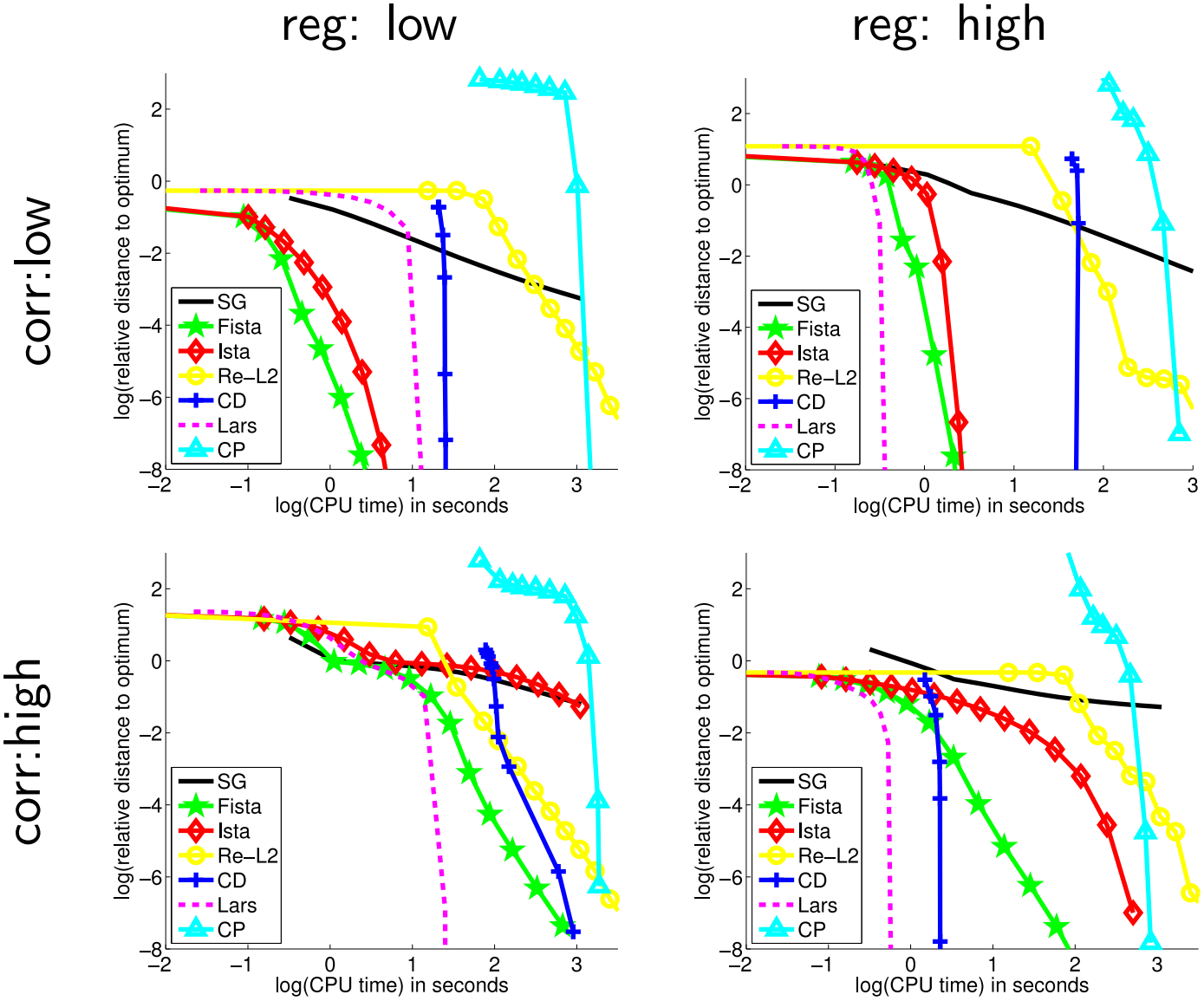
In the worst-case, the regularization path of the Lasso has exactly $(3^p + 1)/2$ *linear segments.*



Lasso Empirical comparison: Lasso, small scale ($n = 200, p = 200$)



Empirical comparison: Lasso, medium scale ($n = 2000, p = 10000$)



Empirical comparison: conclusions

Lasso

- Generic methods (subgradient descent, QP/CP solvers) are slow;
- homotopy fastest in **low dimension** and/or for **high correlation**
- Proximal methods are competitive
  - esp. larger setting and/or weak corr. and/or weak reg. and/or low precision
- Coordinate descent
  - usually dominated by LARS;
  - but much simpler to implement!

Smooth Losses and other regularization

- LARS not available $\rightarrow$ (block) coordinate descent, proximal gradient methods are good candidates.

Conclusion of the part

What was not covered in this part:

- other penalty functions, fused lasso (total variation), hierarchical penalties, structured sparsity with overlapping groups.
- stochastic optimization, Augmented Lagrangian, convergence rates, parallel computing, duality gaps via Fenchel duality...

Software. Try it yourself

`http://www.di.ens.fr/willow/SPAMS/` (Matlab/R/Python/C++).

More reading on optimization

- Bach, Jenatton, Mairal and Obozinski. Optimization with Sparsity-Inducing Penalties. 2012.
- convex optimization: [Boyd and Vandenberghe, 2004, Borwein and Lewis, 2006, Nocedal and Wright, 2006, Bertsekas, 1999, Nesterov, 2004].

Part III: Non-convex Optimization for Sparse Estimation

Greedy Algorithms

Several equivalent non-convex NP-hard problems:

$$\min_{\beta \in \mathbb{R}^p} \underbrace{\frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2}_{\text{residual } \mathbf{r}} + \underbrace{\lambda\|\beta\|_0}_{\text{regularization}},$$

$$\min_{\beta \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \quad \text{s.t.} \quad \|\beta\|_0 \leq L,$$

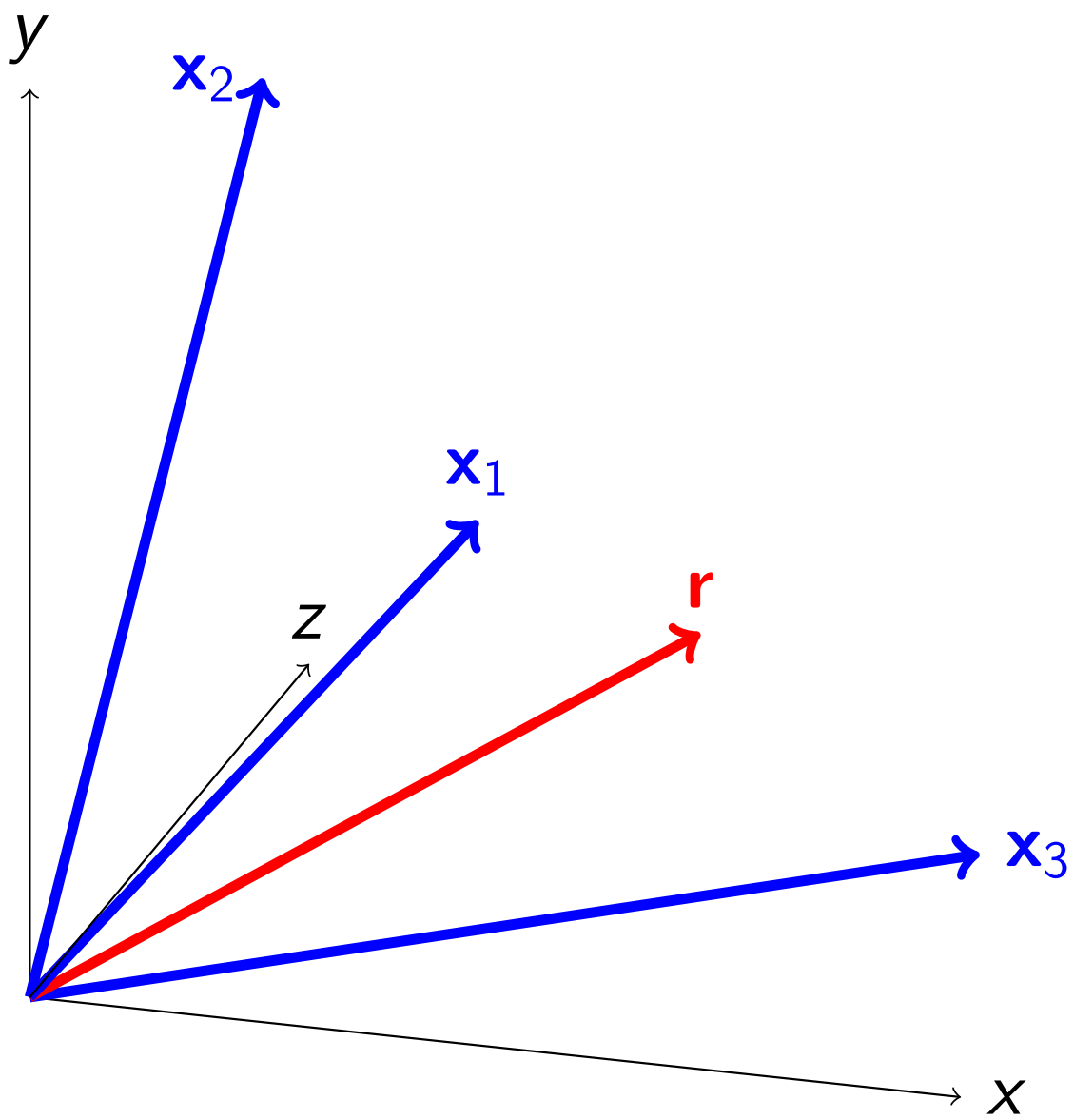
$$\min_{\beta \in \mathbb{R}^p} \|\beta\|_0 \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \leq \varepsilon,$$

The solution is often approximated with a **greedy** algorithm.

- **Signal processing**: Matching Pursuit [Mallat and Zhang, 1993], Orthogonal Matching Pursuit [Pati et al., 1993];
- **Statistics**: L2-boosting [Bühlmann and Yu, 2003], forward selection.

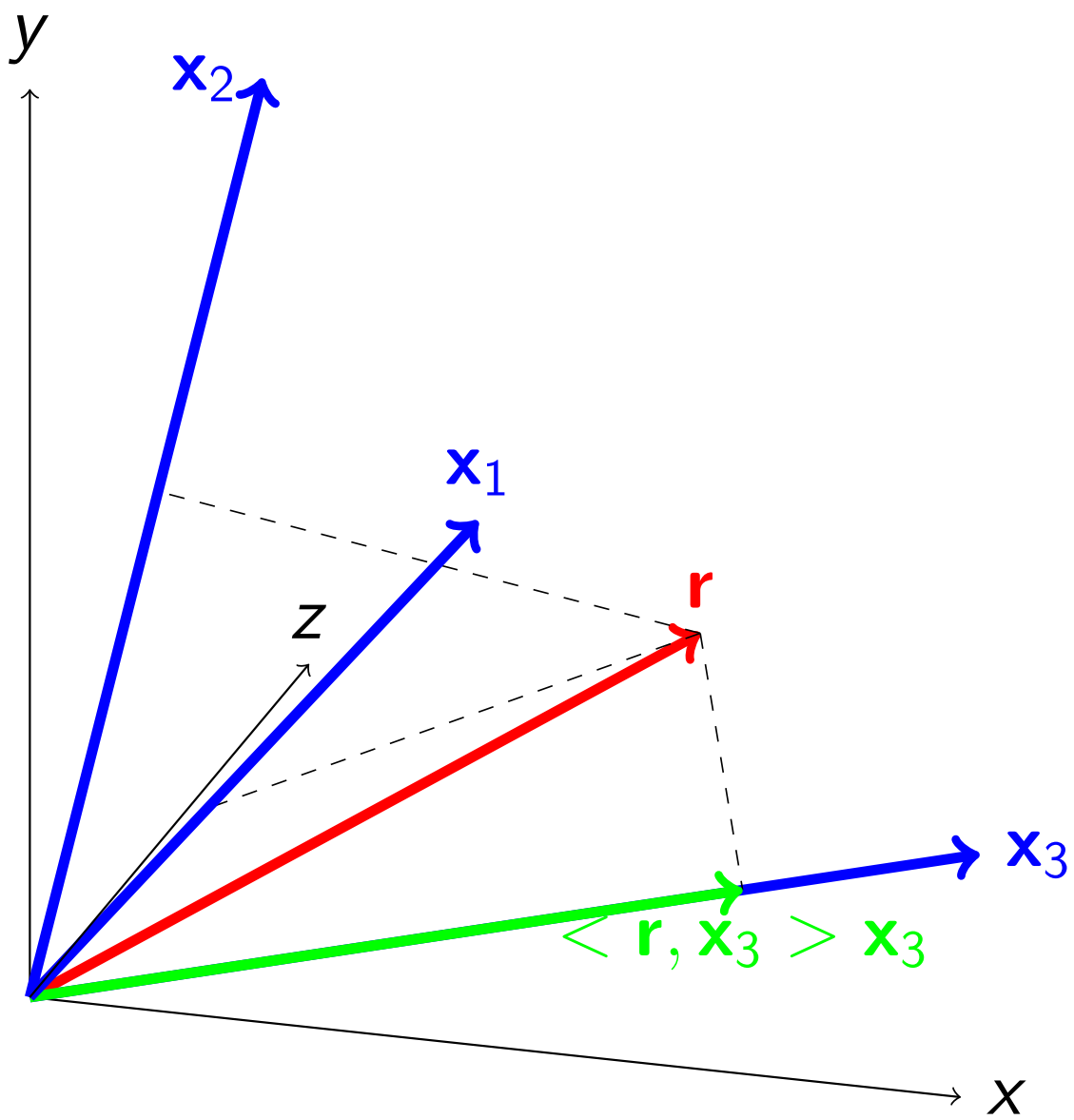
Matching Pursuit

$$\beta = (0, 0, 0)$$



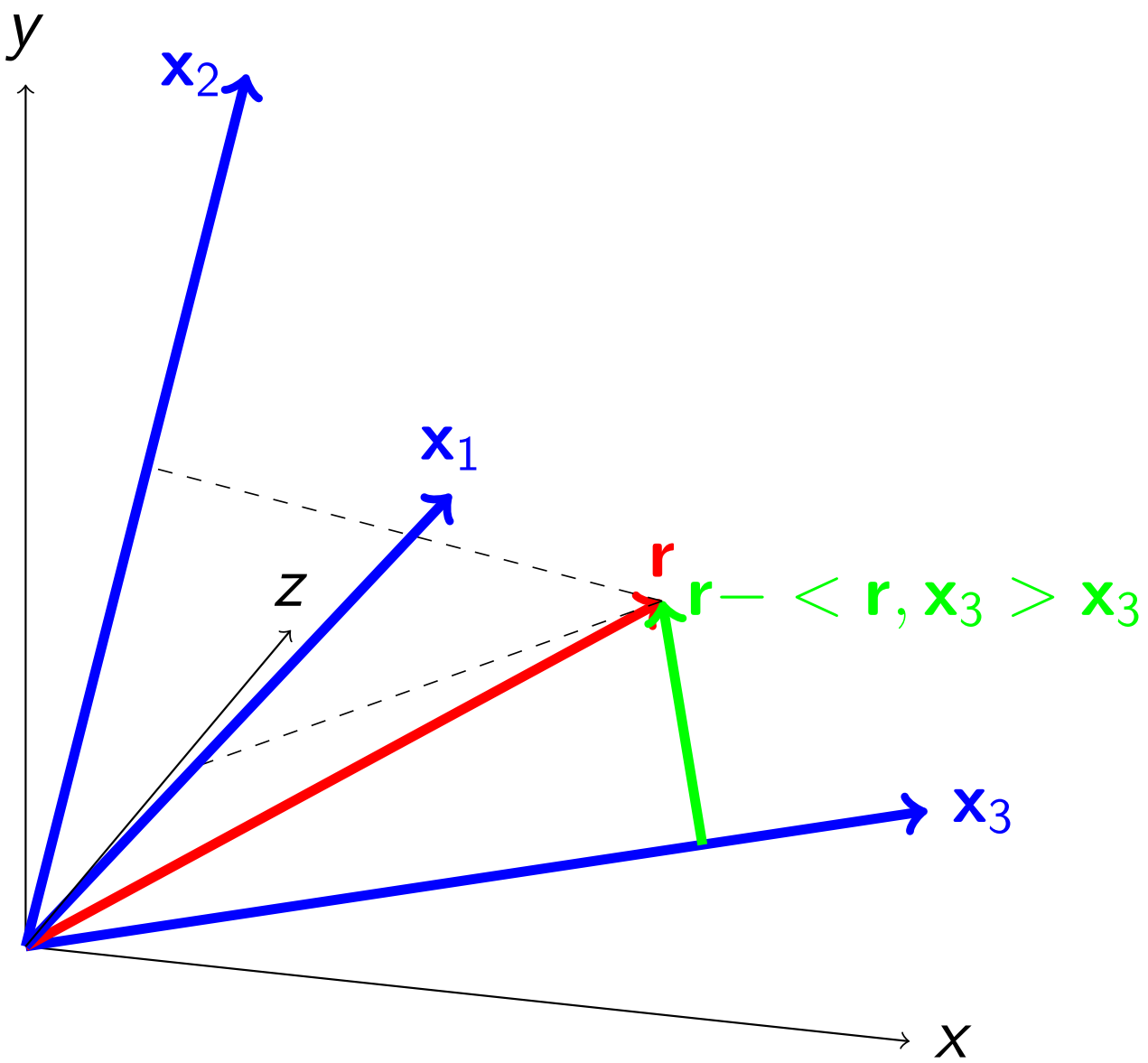
Matching Pursuit

$$\boldsymbol{\beta} = (0, 0, 0)$$



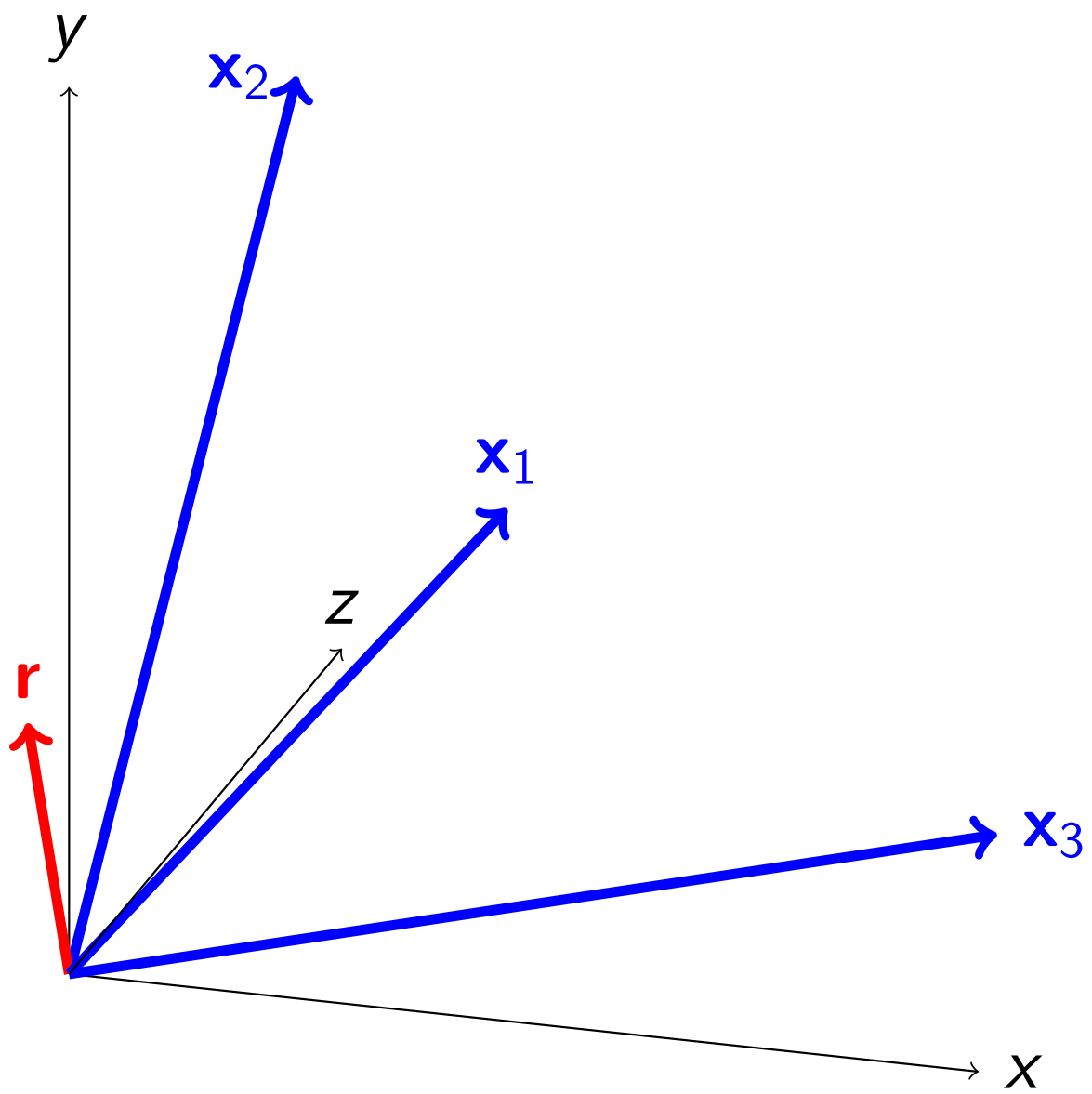
Matching Pursuit

$$\beta = (0, 0, 0)$$



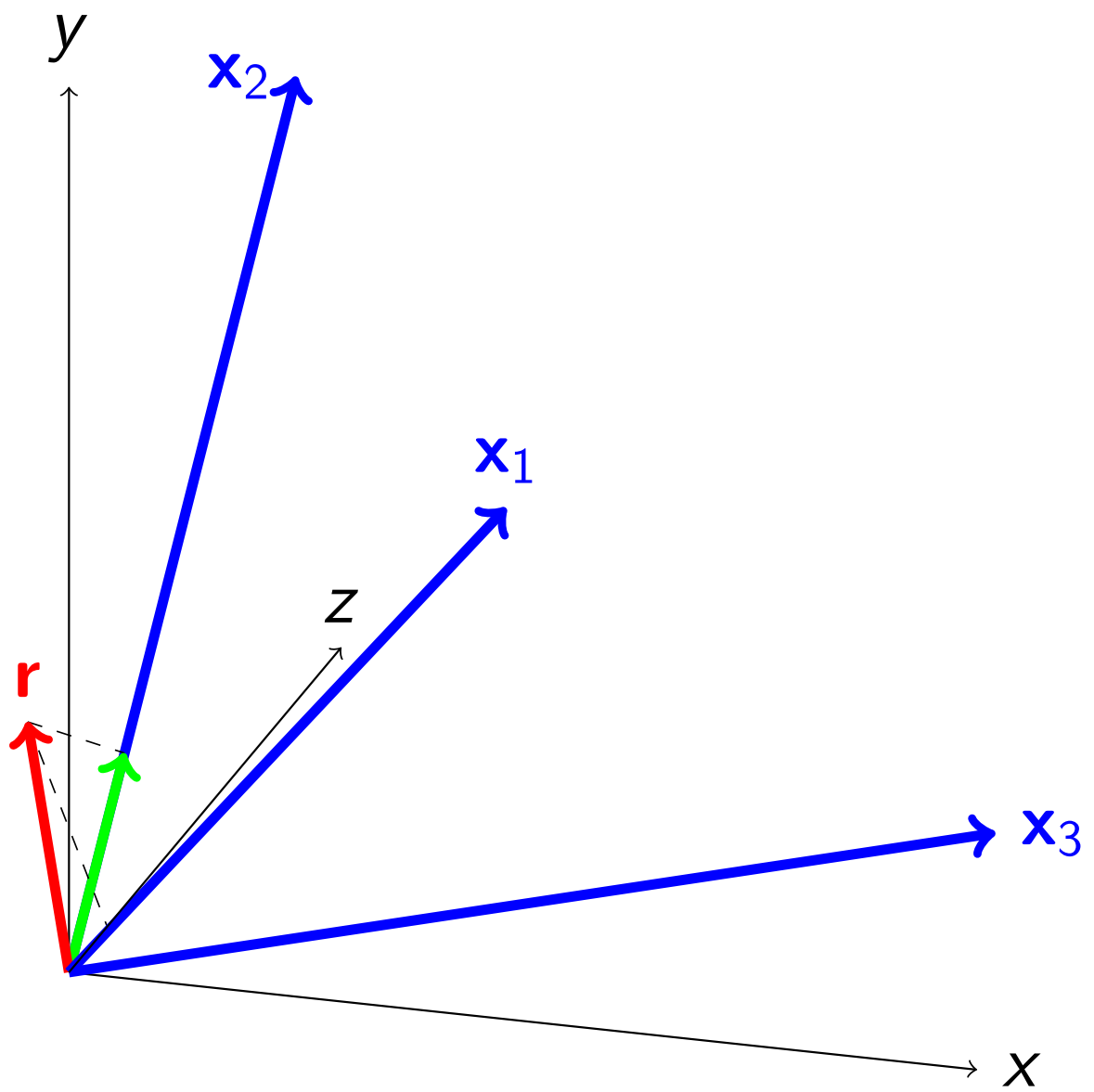
Matching Pursuit

$$\beta = (0, 0, 0.75)$$



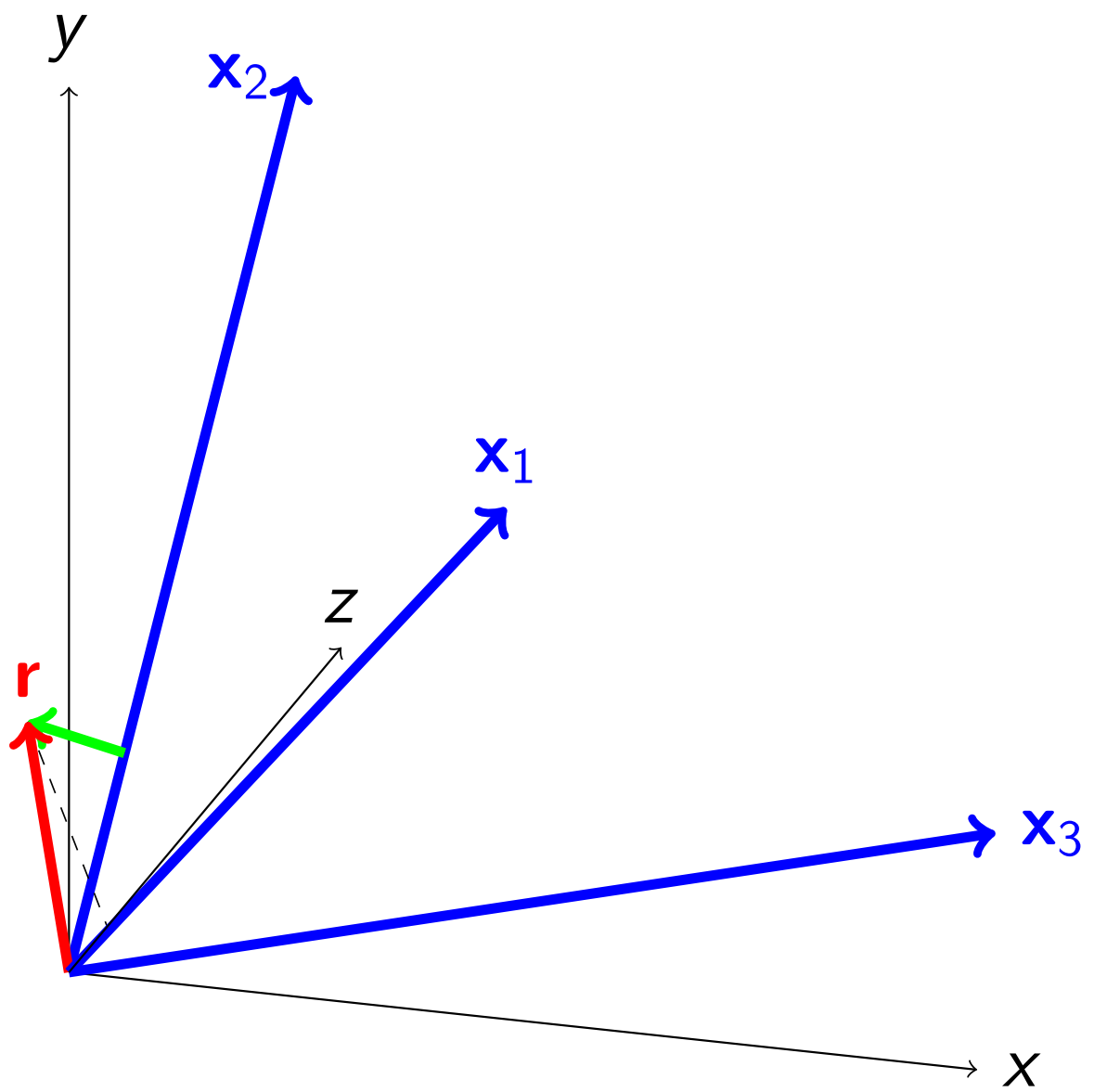
Matching Pursuit

$$\beta = (0, 0, 0.75)$$



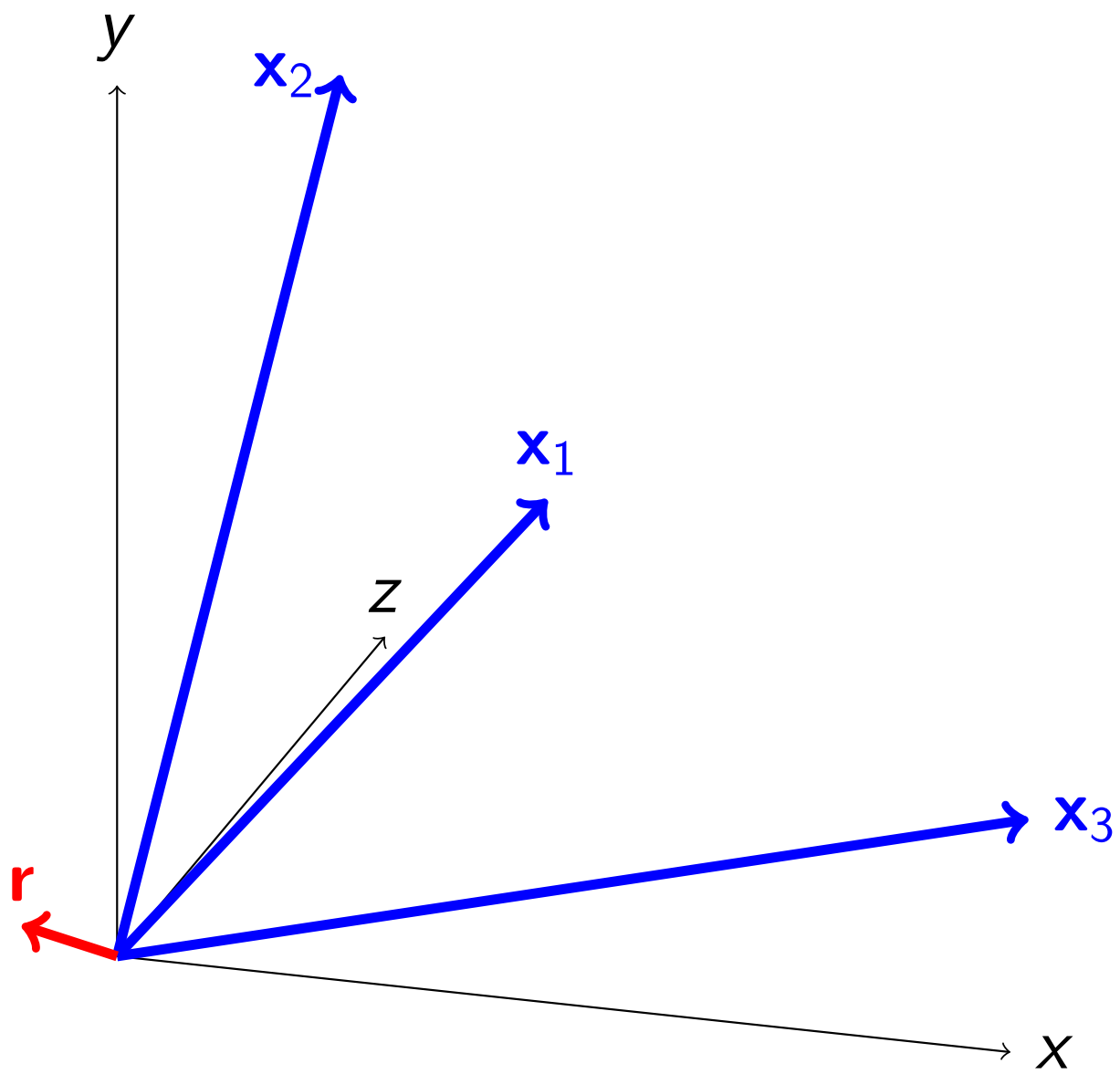
Matching Pursuit

$$\beta = (0, 0, 0.75)$$



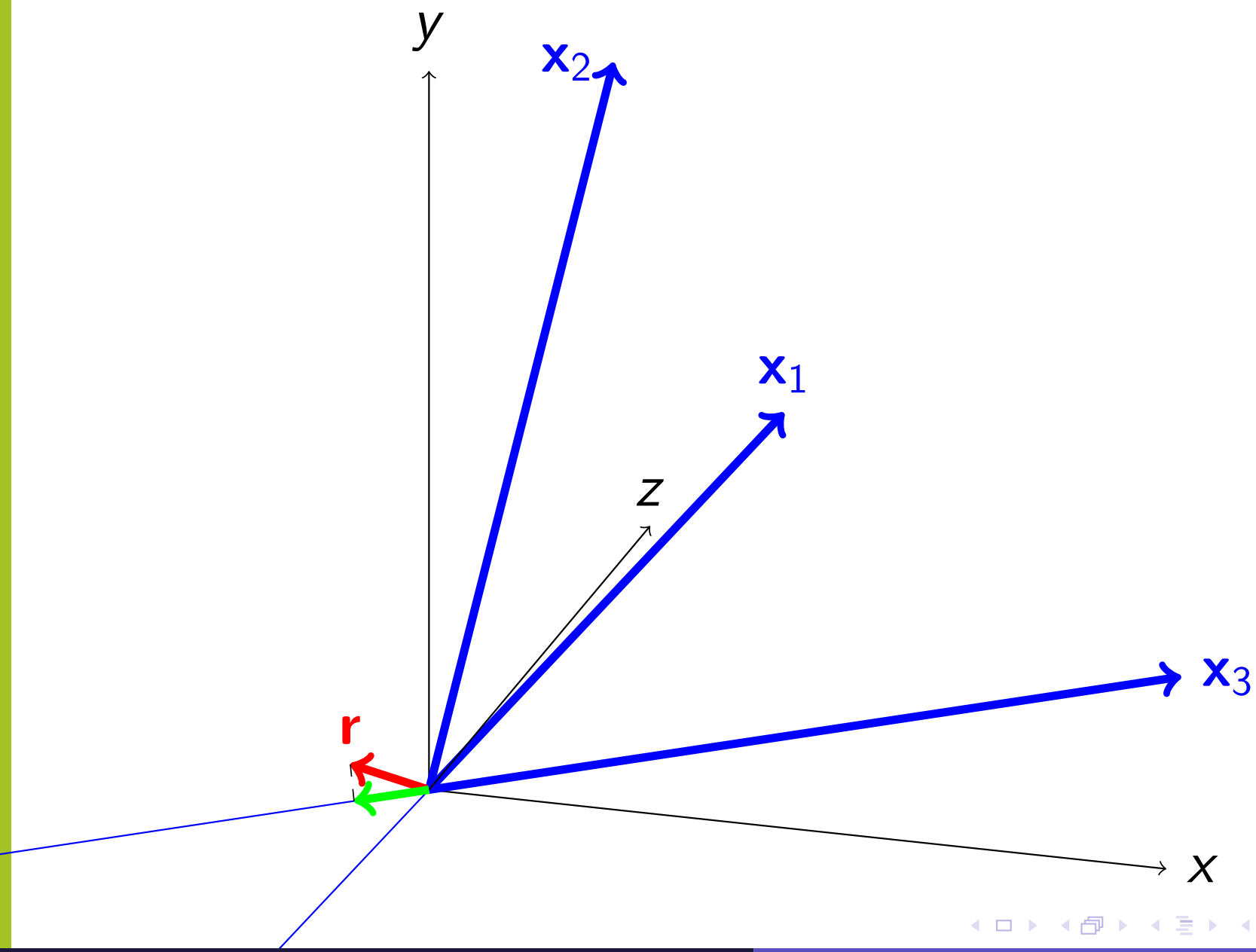
Matching Pursuit

$$\beta = (0, 0.24, 0.75)$$



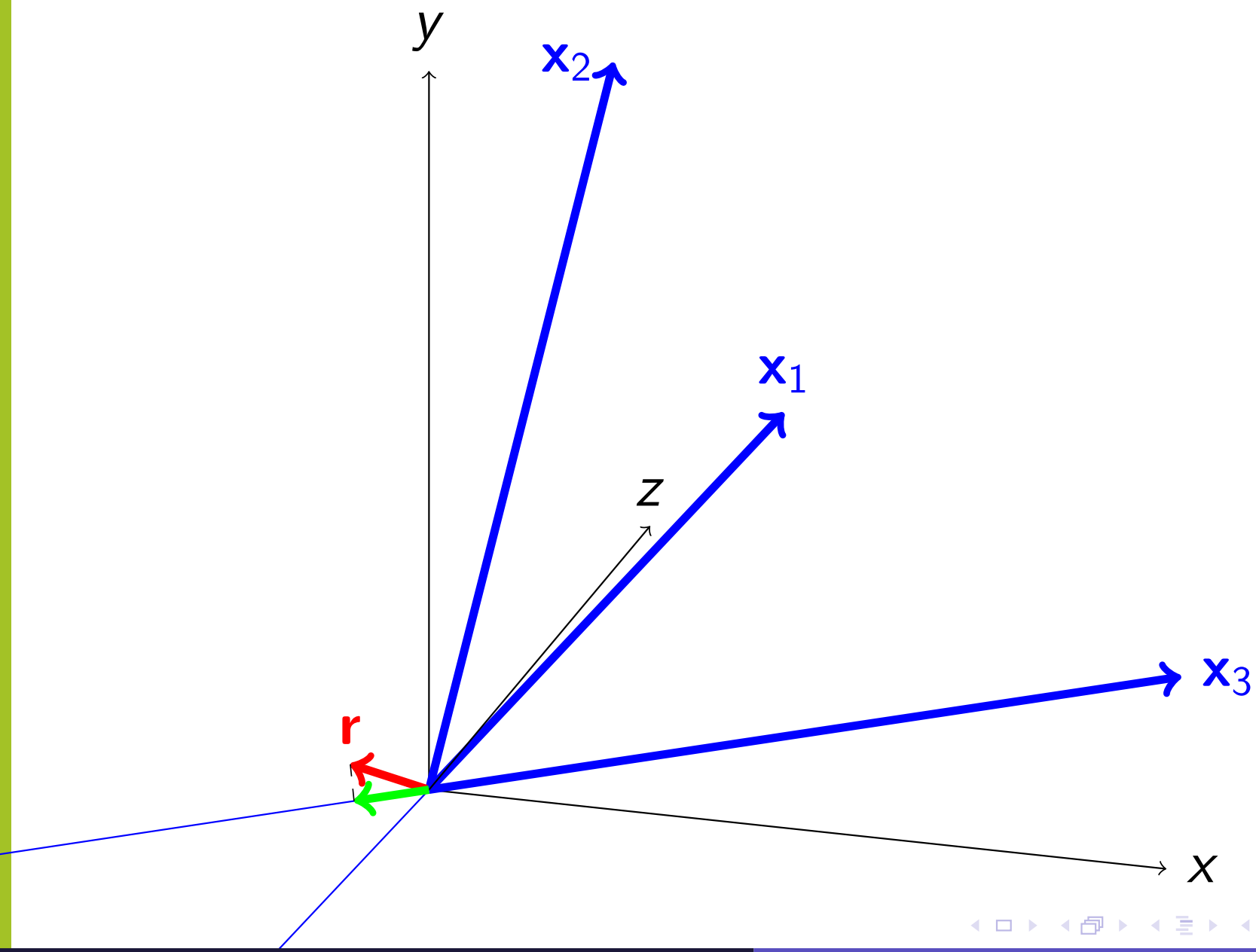
Matching Pursuit

$$\beta = (0, 0.24, 0.75)$$



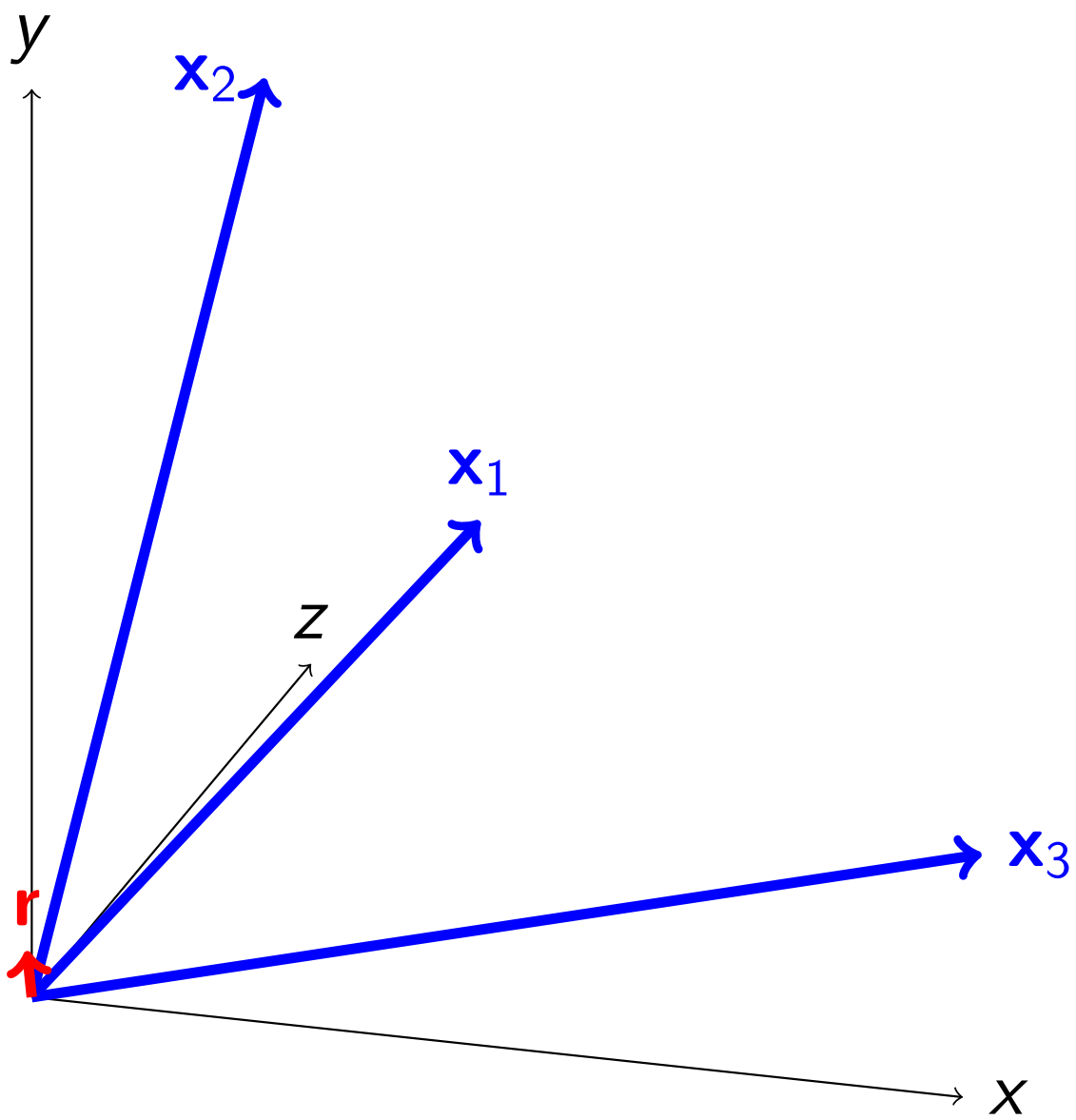
Matching Pursuit

$$\beta = (0, 0.24, 0.75)$$



Matching Pursuit

$$\beta = (0, 0.24, 0.65)$$



Matching Pursuit

$$\min_{\beta \in \mathbb{R}^p} \| \underbrace{\mathbf{y} - \mathbf{X}\beta}_{\mathbf{r}} \|_2^2 \quad \text{s.t.} \quad \|\beta\|_0 \leq L$$

1: $\beta \leftarrow 0$
2: $\mathbf{r} \leftarrow \mathbf{y}$ (residual).
3: **while** $\|\beta\|_0 < L$ **do**
4: Select the predictor with maximum correlation with the residual

$$\hat{\jmath} \leftarrow \arg\max_{j=1,\ldots,p} |\mathbf{x}^{j\top} \mathbf{r}|$$

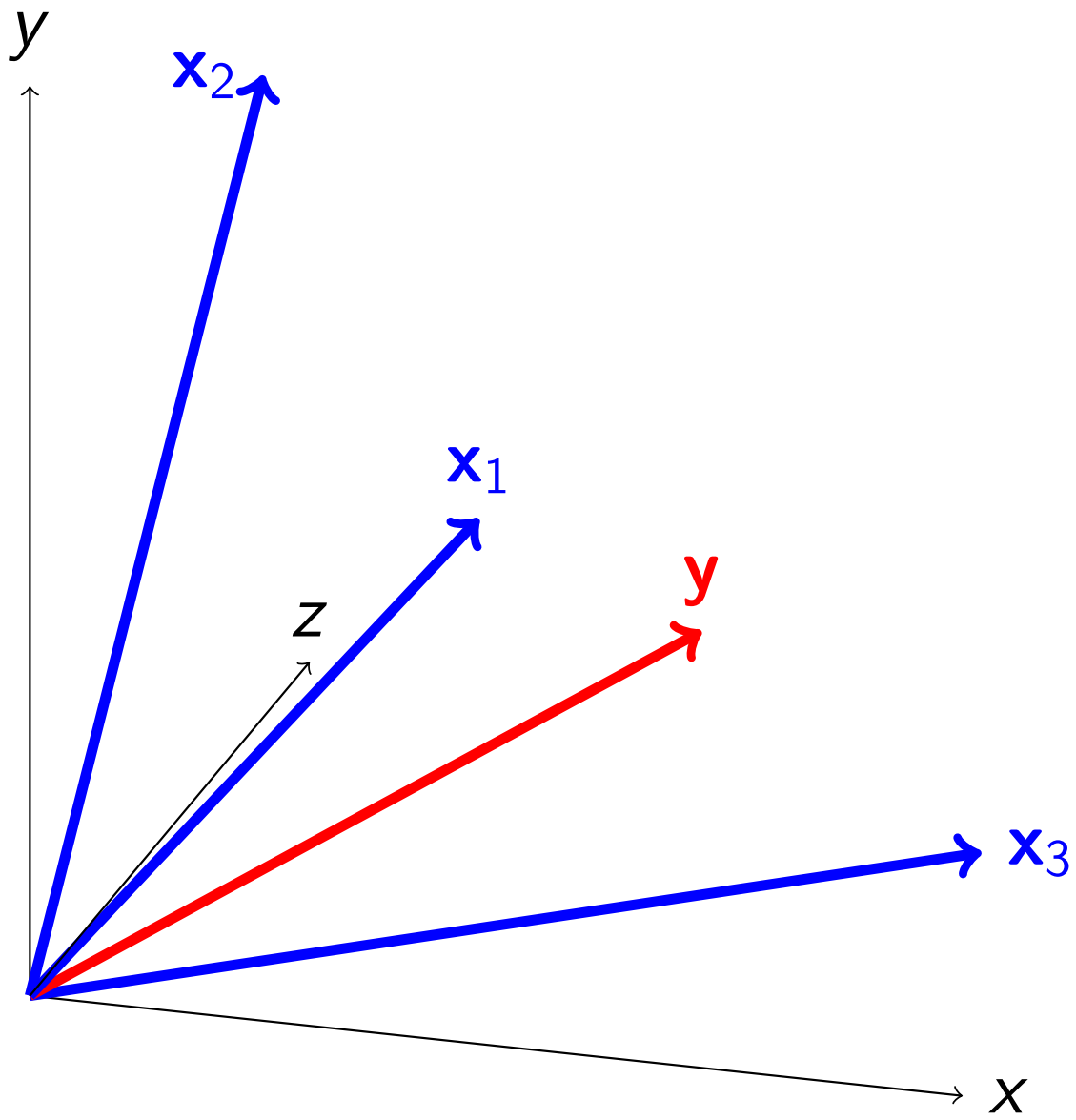
5: Update the residual and the coefficients

$$\begin{aligned} \beta_{\hat{\jmath}} &\leftarrow \beta_{\hat{\jmath}} + \mathbf{x}^{\hat{\jmath}\top} \mathbf{r} \\ \mathbf{r} &\leftarrow \mathbf{r} - (\mathbf{x}^{\hat{\jmath}\top} \mathbf{r}) \mathbf{x}^{\hat{\jmath}} \end{aligned}$$

6: **end while**

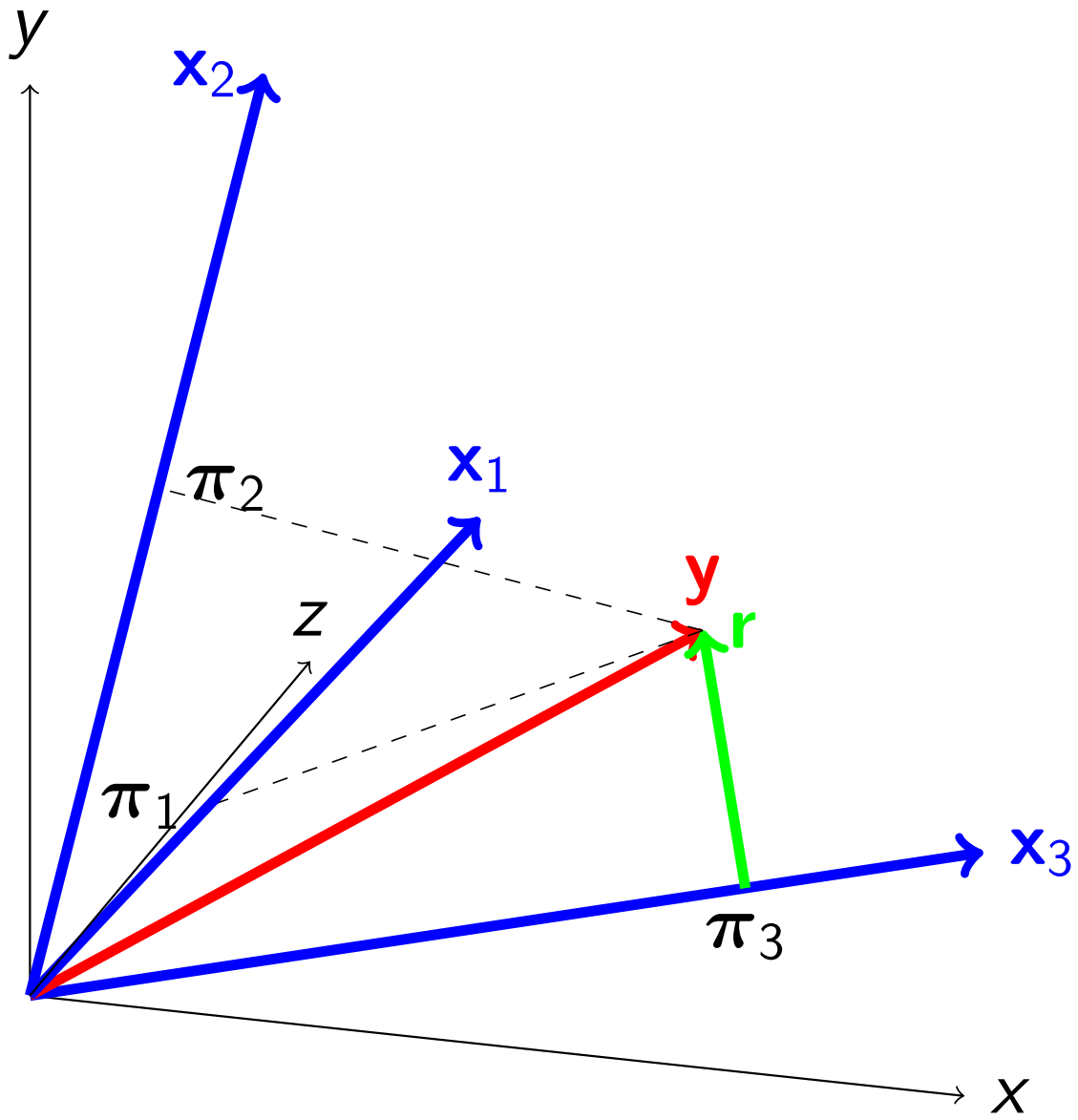
Orthogonal Matching Pursuit

$$\boldsymbol{\beta} = (0, 0, 0)$$
$$\Gamma = \emptyset$$



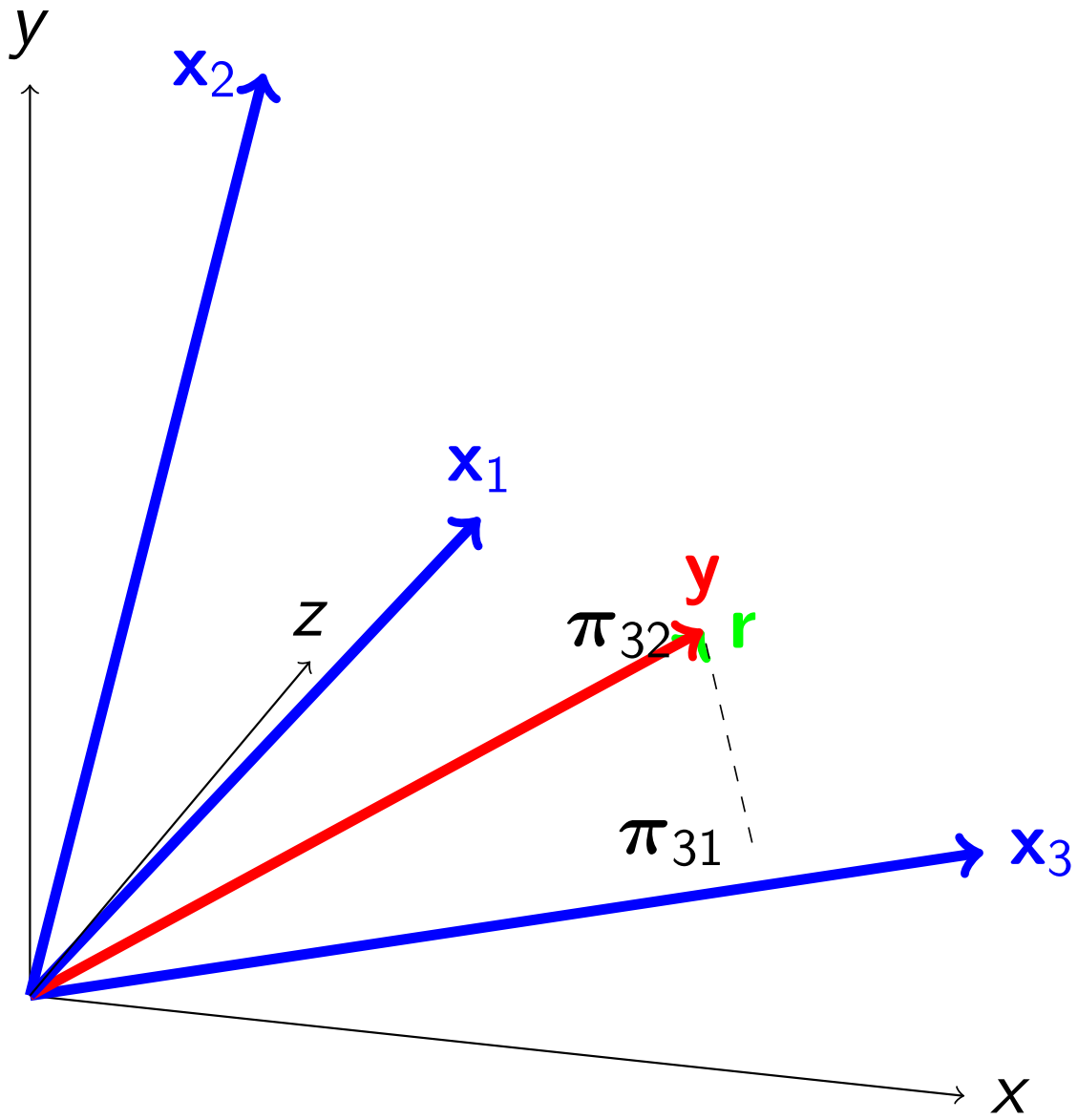
Orthogonal Matching Pursuit

$$\beta = (0, 0, 0.75)$$
$$\Gamma = \{3\}$$



Orthogonal Matching Pursuit

$$\beta = (0, 0.29, 0.63)$$
$$\Gamma = \{3, 2\}$$



Orthogonal Matching Pursuit

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leq L$$

1: $\Gamma = \emptyset$.
2: **for** *iter* $= 1, \ldots, L$ **do**
3: Select the predictor which most reduces the objective

$$\hat{\imath} \leftarrow \arg\min_{i \in \Gamma^c} \left\{ \min_{\boldsymbol{\beta}'} \|\mathbf{y} - \mathbf{X}_{\Gamma \cup \{i\}} \boldsymbol{\beta}'\|_2^2 \right\}$$

4: Update the active set: $\Gamma \leftarrow \Gamma \cup \{\hat{\imath}\}$.
5: Update the residual (orthogonal projection)

$$\mathbf{r} \leftarrow (\mathbf{I} - \mathbf{X}_\Gamma (\mathbf{X}_\Gamma^\top \mathbf{X}_\Gamma)^{-1} \mathbf{X}_\Gamma^\top)\mathbf{y}.$$

6: Update the coefficients

$$\boldsymbol{\beta}_\Gamma \leftarrow (\mathbf{X}_\Gamma^\top \mathbf{X}_\Gamma)^{-1} \mathbf{X}_\Gamma^\top \mathbf{y}.$$

7: **end for**

Orthogonal Matching Pursuit

The keys for a good implementation

- If available, use Gram matrix $\mathbf{G} = \mathbf{X}^\top \mathbf{X}$,
- Maintain the computation of $\mathbf{X}^\top \mathbf{r}$ for each signal,
- Maintain a Cholesky decomposition of $(\mathbf{X}_\Gamma^\top \mathbf{X}_\Gamma)^{-1}$ for each signal.

The total complexity for decomposing n L-sparse signals of size m with a dictionary of size p is

$$\underbrace{O(p^2 m)}_{\text{Gram matrix}} + \underbrace{O(nL^3)}_{\text{Cholesky}} + \underbrace{O(n(pm + pL^2))}_{\mathbf{X}^T\mathbf{r}} = O(np(m + L^2))$$

It is also possible to use the matrix inversion lemma instead of a Cholesky decomposition (same complexity, but less numerical stability).

Example with the software SPAMS

Software available at `http://www.di.ens.fr/willow/SPAMS/`

```
>> I=double(imread('data/lena.eps'))/255;
>> %extract all patches of I
>> X=im2col(I,[8 8],'sliding');
>> %load a dictionary of size 64 x 256
>> D=load('dict.mat');
>>
>> %set the sparsity parameter L to 10
>> param.L=10;
>> alpha=mexOMP(X,D,param);
```

On a 4-cores 3.6Ghz machine: **150000 signals processed per second!**

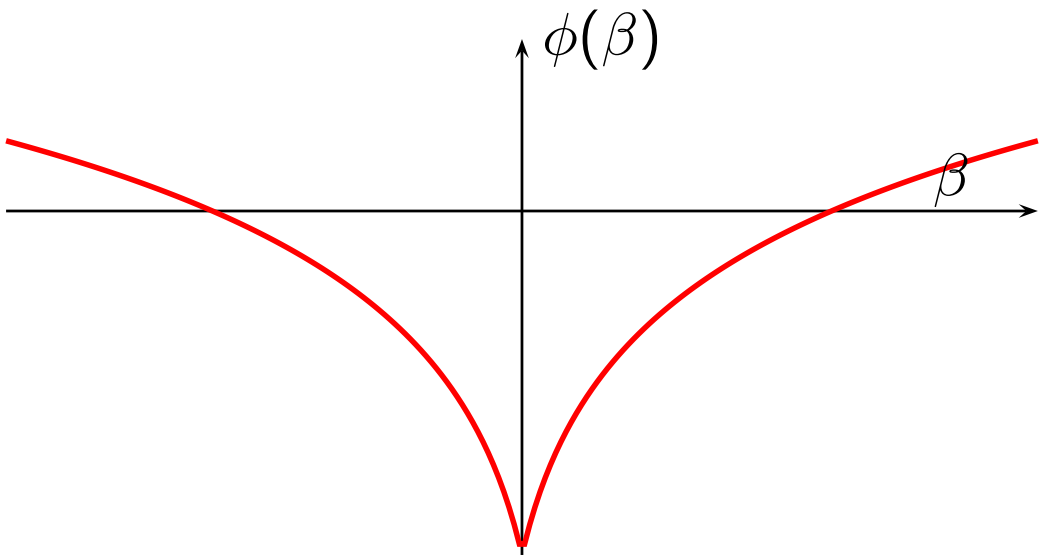
DC (difference of convex) - Programming

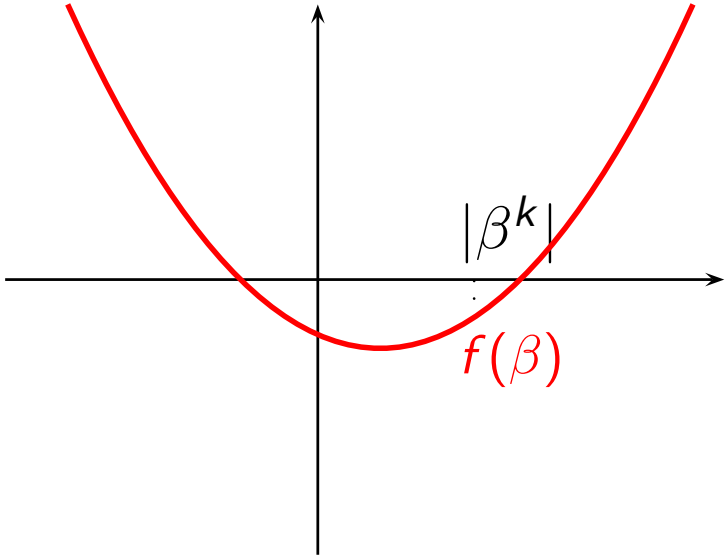
Remember? Concave functions with a kink at zero

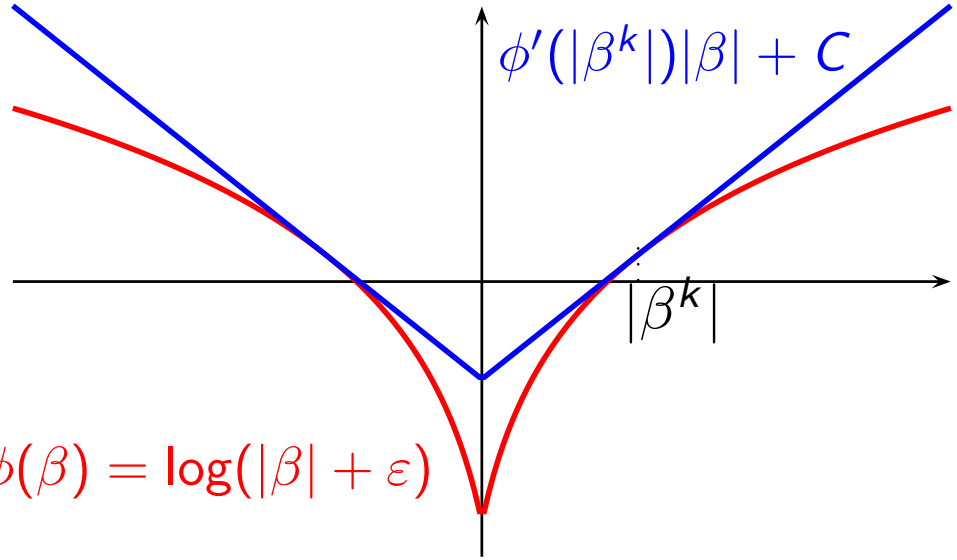
$\Omega(\boldsymbol{\beta}) = \sum_{j=1}^{p} \phi(|\boldsymbol{\beta}_j|)$.

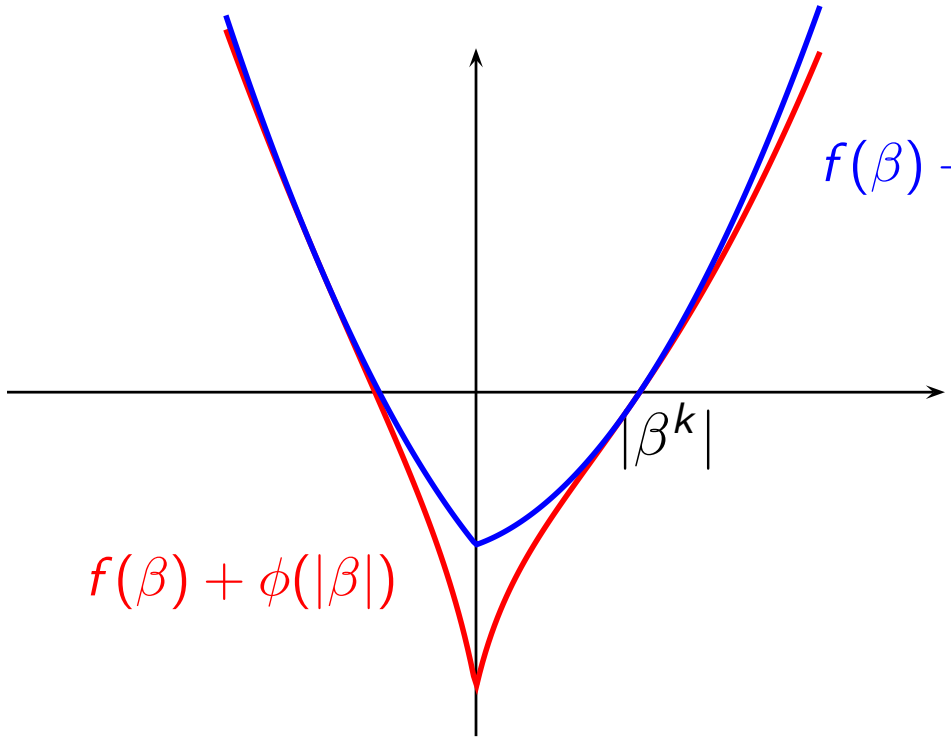
- ℓ_q-"pseudo-norm", with $0 < q < 1$: $\Omega(\boldsymbol{\beta}) \triangleq \sum_{j=1}^{p} (|\boldsymbol{\beta}_j| + \varepsilon)^q$,
- log penalty, $\Omega(\boldsymbol{\beta}) \triangleq \sum_{j=1}^{p} \log(|\boldsymbol{\beta}_j| + \varepsilon)$,

ϕ is any function that looks like this:




$|\beta^k|$
$f(\beta)$


$\phi'(|\beta^k|)|\beta| + C$
$|\beta^k|$
$\phi(\beta) = \log(|\beta| + \varepsilon)$


$f(\beta) + \phi'(|\beta^k|)|\beta| + C$
$|\beta^k|$
$f(\beta) + \phi(|\beta|)$

DC (difference of convex) - Programming

$$\min_{\beta \in \mathbb{R}^p} f(\beta) + \lambda \sum_{j=1}^{p} \phi(|\beta_j|).$$

This problem is non-convex. f is convex, and ϕ is concave on $\mathbb{R}^+$.
if β^k is the current estimate at iteration k, the algorithm solves

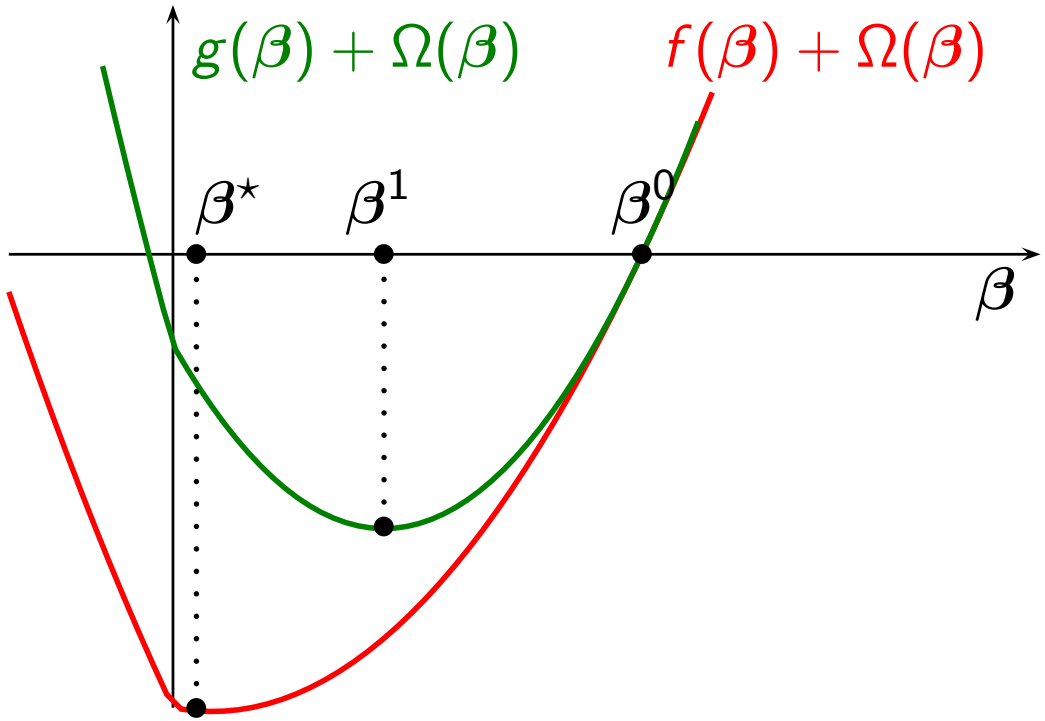
$$\beta^{k+1} \leftarrow \arg\min_{\beta \in \mathbb{R}^p} \left[f(\beta) + \lambda \sum_{j=1}^{p} \psi'(|\beta_j^k|)|\beta_j| \right],$$

which is a **reweighted**-ℓ_1 problem.
Warning: It does not solve the non-convex problem, only provides a stationary point.
In practice, each iteration sets to zero small coefficients. After $2-3$ iterations, the result does not change much.

The Proximal Gradient Method

This inequality is also true for non-convex functions!

If ∇f is L-Lipschitz continuous



- $f(\beta) + \Omega(\beta) \leq f(\beta^0) + \nabla f(\beta^0)^\top(\beta - \beta^0) + \frac{L}{2}\|\beta - \beta^0\|_2^2 + \Omega(\beta)$;

The Proximal Gradient Method

As before, the method consits of the following iteration

$$\boxed{\beta^t \leftarrow \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \left\| \beta^{t-1} - \frac{1}{L} \nabla f(\beta^{t-1}) - \beta \right\|_2^2 + \Omega(\beta).}$$

- It requires computing efficiently the **proximal operator** of Ω.

$$\alpha \mapsto \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|\alpha - \beta\|_2^2 + \Omega(\beta).$$

- For the ℓ_0-penalty ($\Omega(\beta) = \lambda \|\beta\|_0$), this amounts to a hard-thresholding:

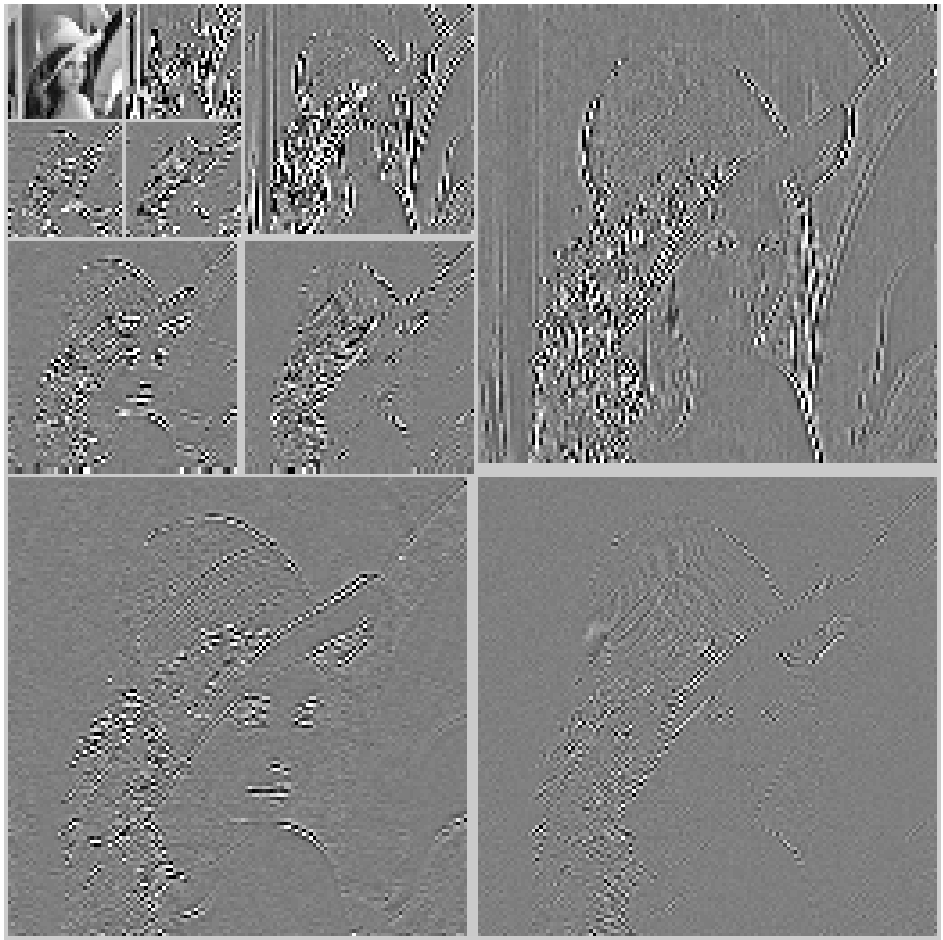
$$\beta_j^\star = 1_{|\alpha_j| \geq \sqrt{2\lambda}} \alpha_j.$$

Trick: start with a high value for λ, and gradually decrease it.

Part IV: Structured Sparsity

Wavelet coefficients

- Zero-tree wavelets coding [Shapiro, 1993];
- block thresholding [Cai, 1999].



Sparse linear models for natural image patches

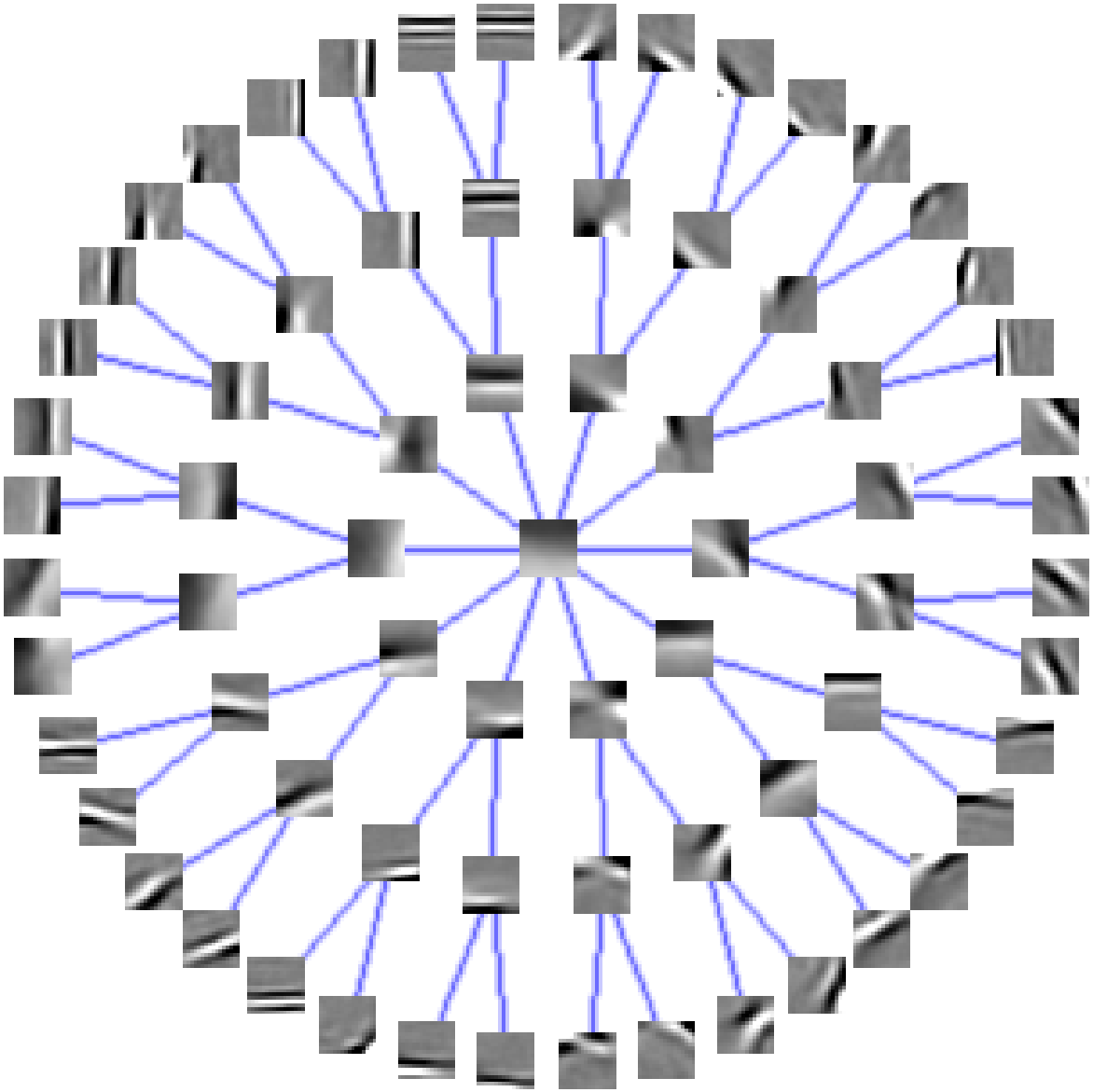
Image restoration

Sparse linear models for natural image patches

Image restoration

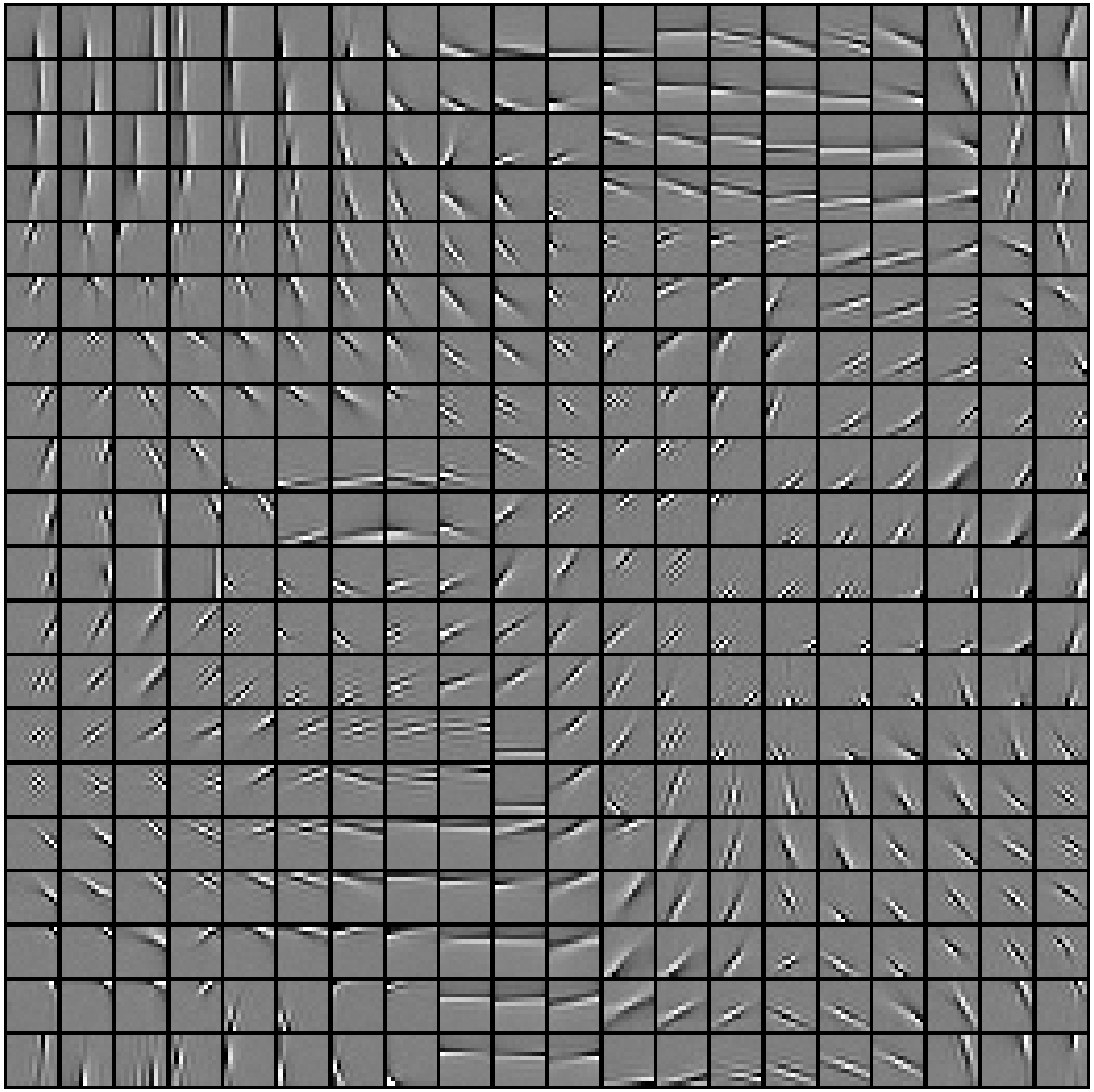
Structured dictionary for natural image patches

[Jenatton, Mairal, Obozinski, and Bach, 2010]



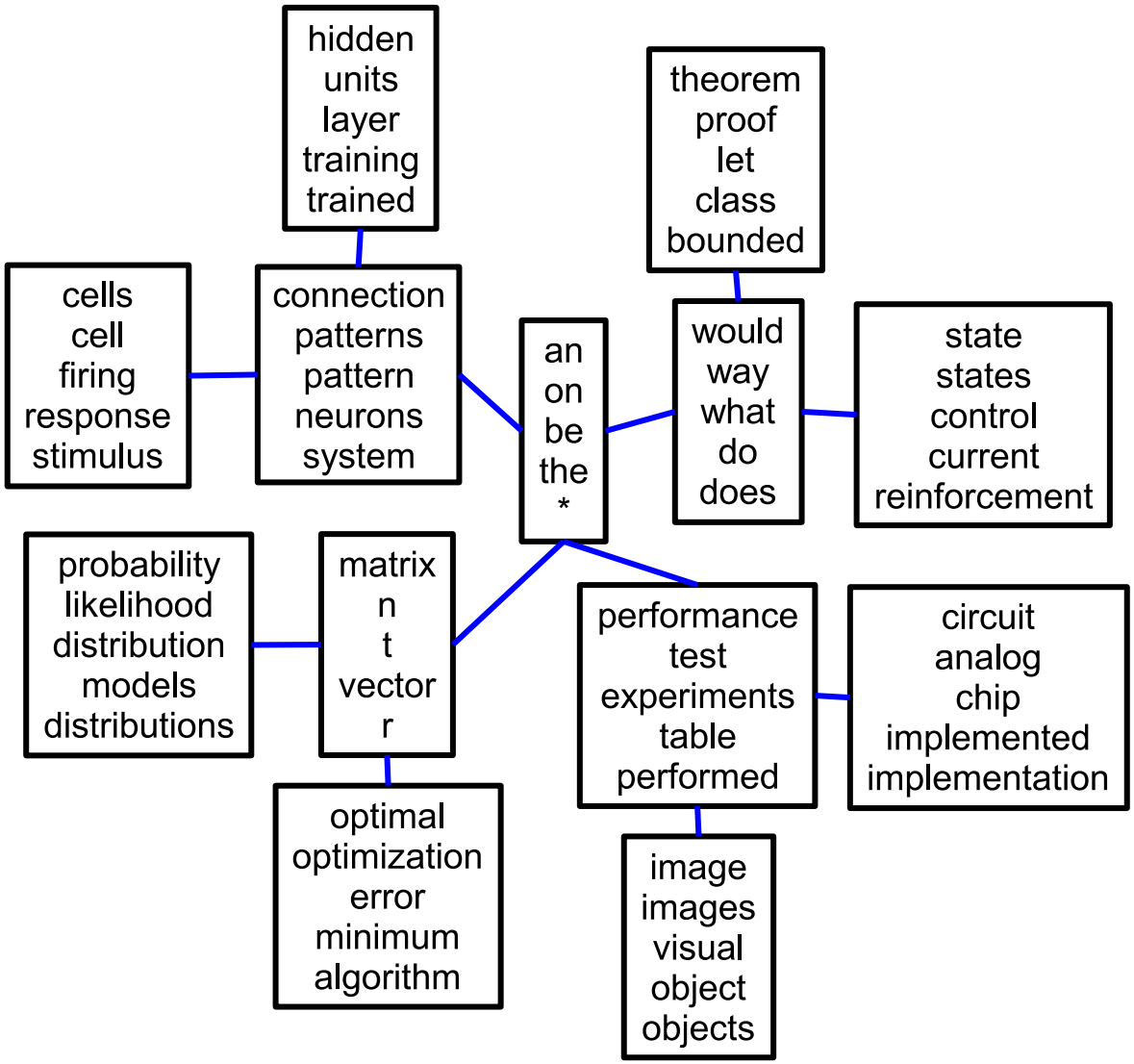
Structured dictionary for natural image patches

[Mairal, Jenatton, Obozinski, and Bach, 2011]



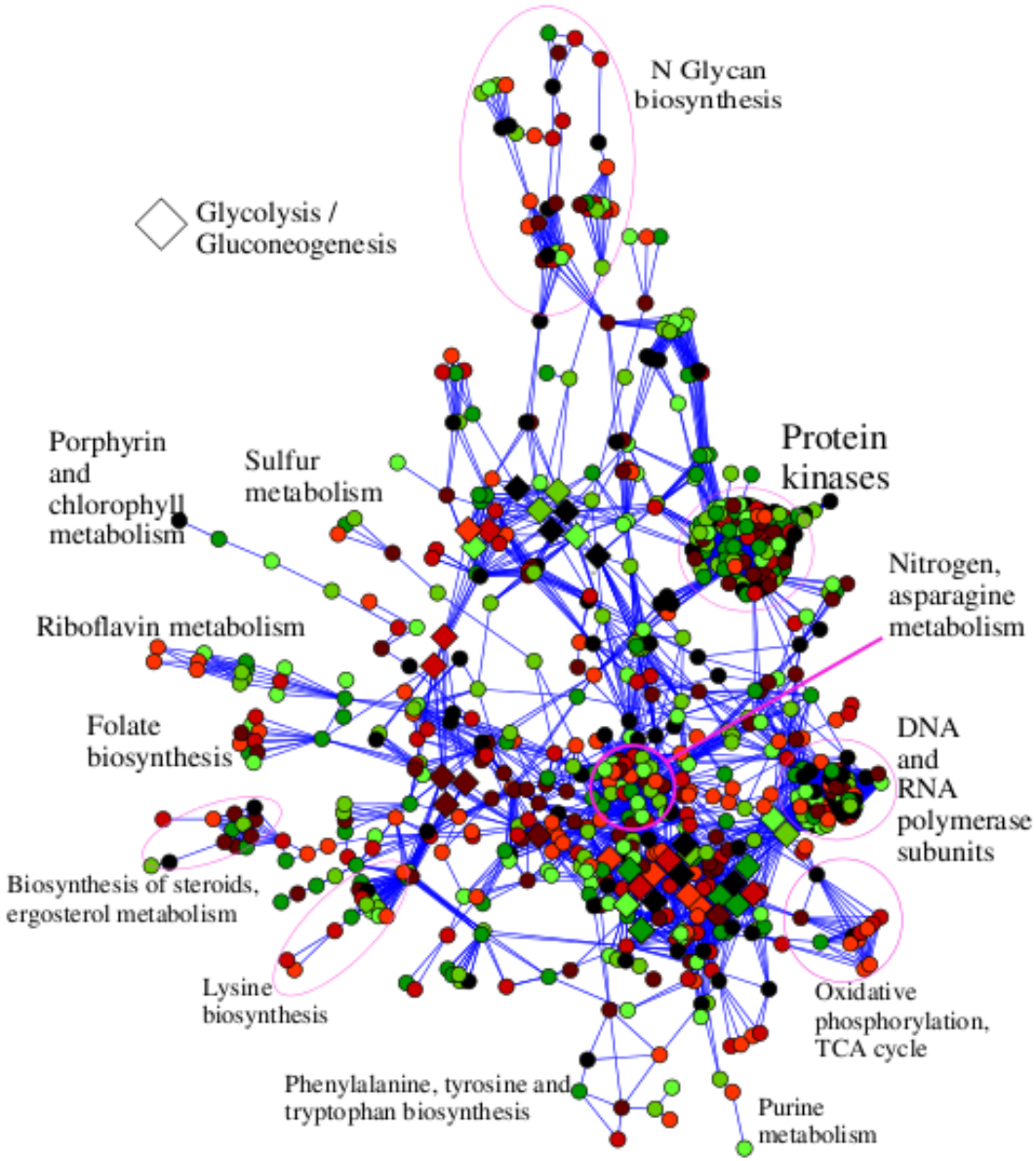
Tree of topics

[Jenatton, Mairal, Obozinski, and Bach, 2010]



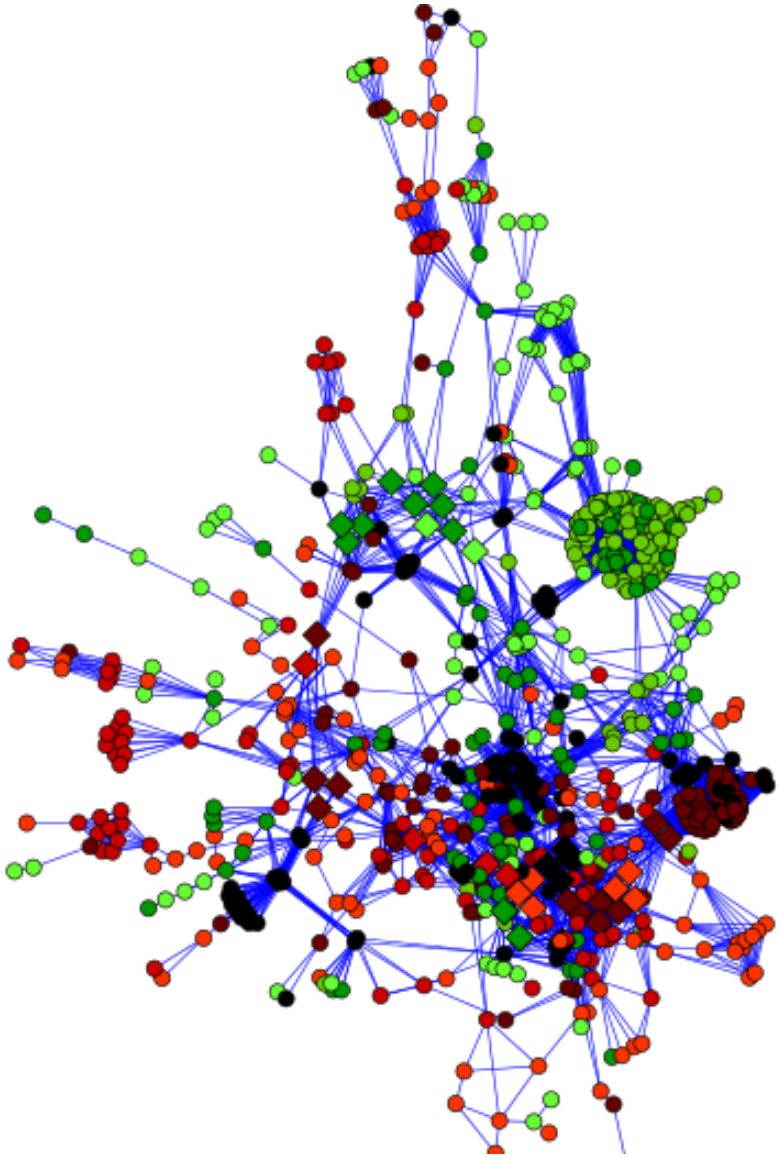
Metabolic network of the budding yeast

from Rapaport, Zinovyev, Dutreix, Barillot, and Vert [2007]



Metabolic network of the budding yeast

from Rapaport, Zinovyev, Dutreix, Barillot, and Vert [2007]



Questions about structured sparsity

$$\min_{\beta \in \mathbb{R}^p} \underbrace{f(\beta)}_{\text{convex, smooth}} + \underbrace{\lambda \Omega(\beta)}_{\text{regularization}},$$

Ω should encode some a priori knowledge about β.

☺ In this part, we will see

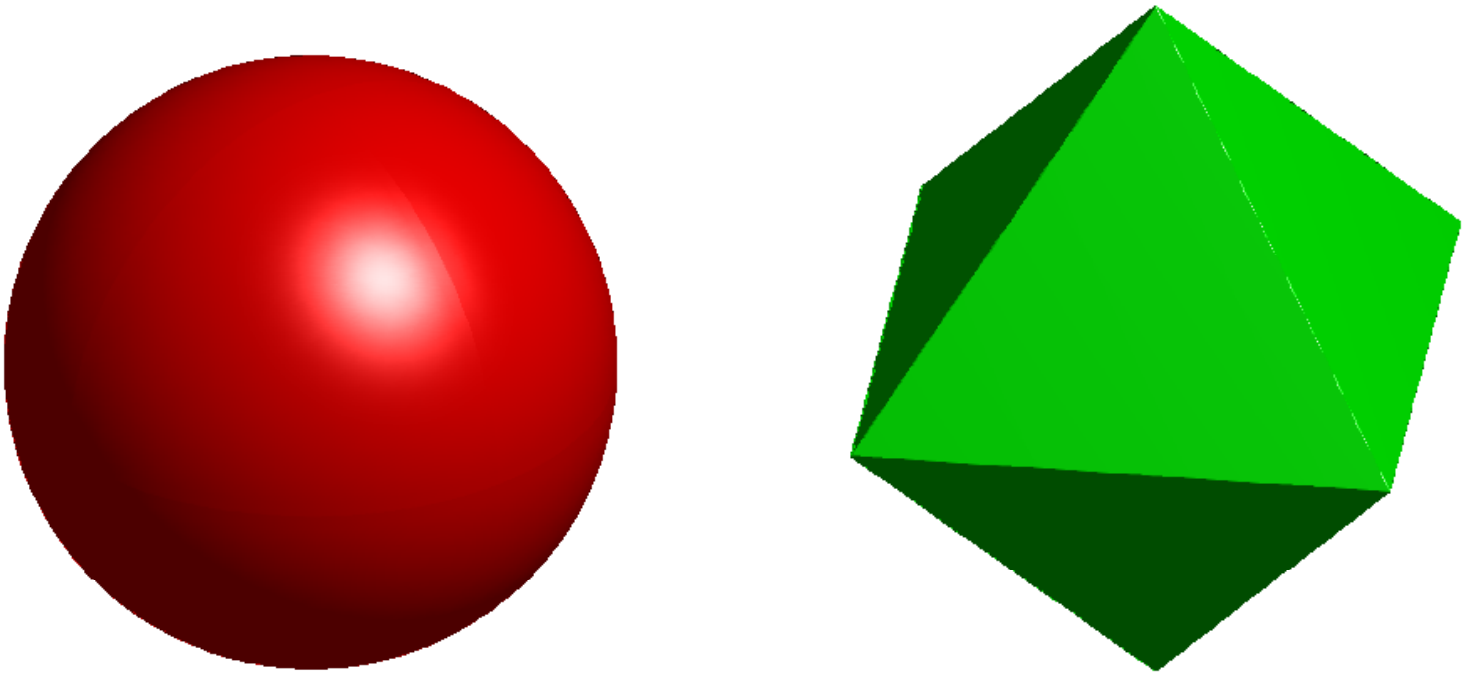
- how to design structured sparsity-inducing functions Ω;
- How to solve the corresponding estimation/inverse problems.

☹ out of the scope of this part:

- consistency, recovery, theoretical properties (statistics...)

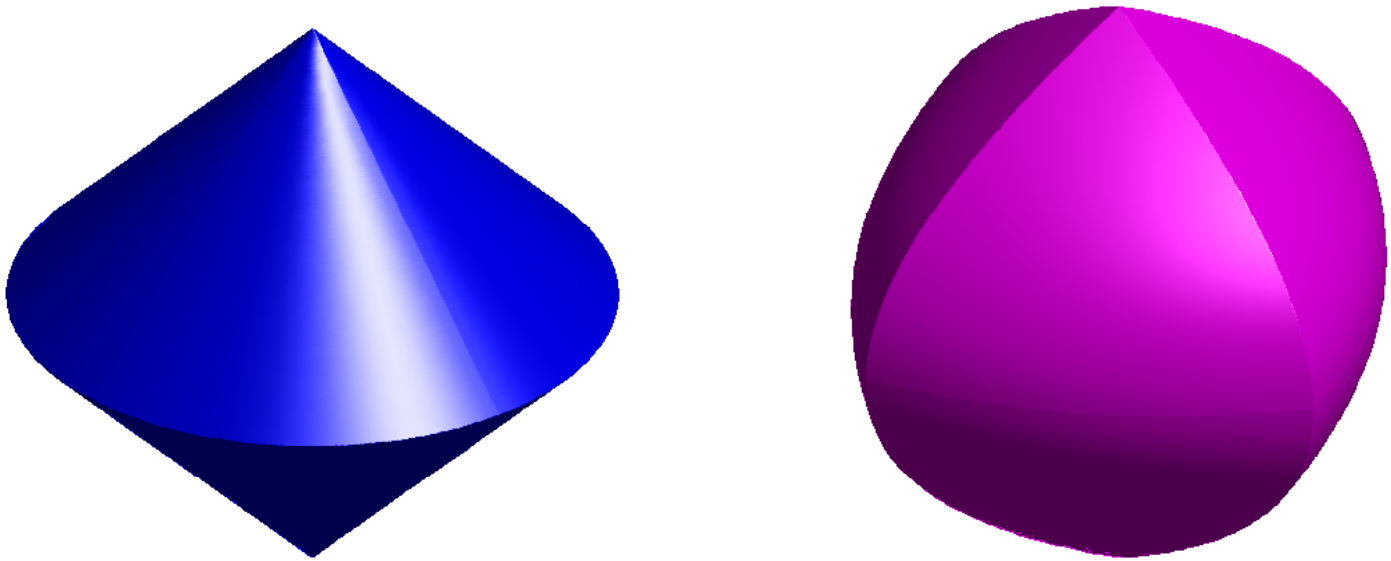
In 3D.

Copyright G. Obozinski



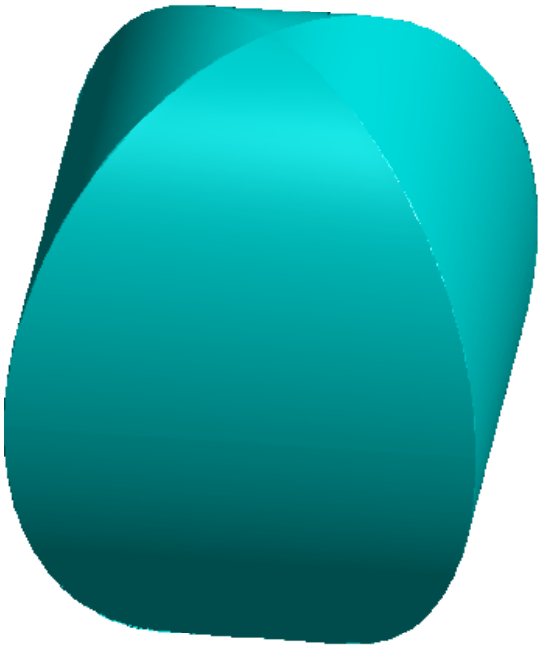
What about more complicated norms?

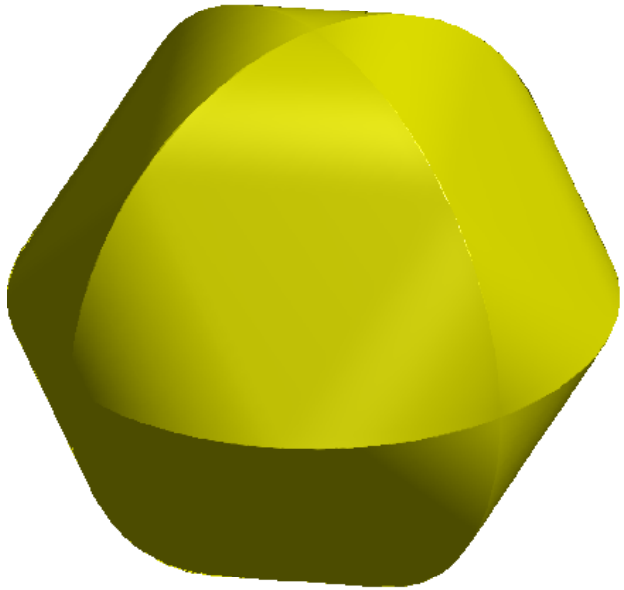
Copyright G. Obozinski



What about more complicated norms?

Copyright G. Obozinski



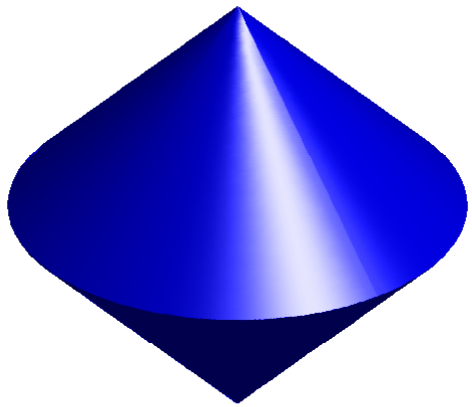


Group Lasso

Turlach et al. [2005], Yuan and Lin [2006], Zhao et al. [2009]

$$\text{the } \ell_1/\ell_q\text{-norm}: \qquad \Omega(\beta) = \sum_{g \in \mathcal{G}} \|\beta_g\|_q.$$

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



$$\Omega(\beta) = \|\beta_{\{1,2\}}\|_2 + |\beta_3|.$$

Structured sparsity with overlapping groups

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., 2009]**;
   - other norms: [Micchelli et al., 2011].

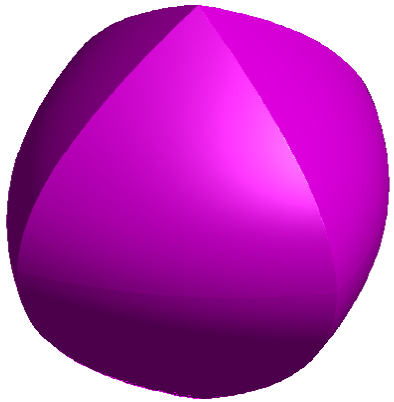
Group Lasso with overlapping groups

[Jenatton, Audibert, and Bach, 2009]

$$\Omega(\beta) = \sum_{g \in \mathcal{G}} \|\beta_g\|_q.$$

What happens when the groups overlap?

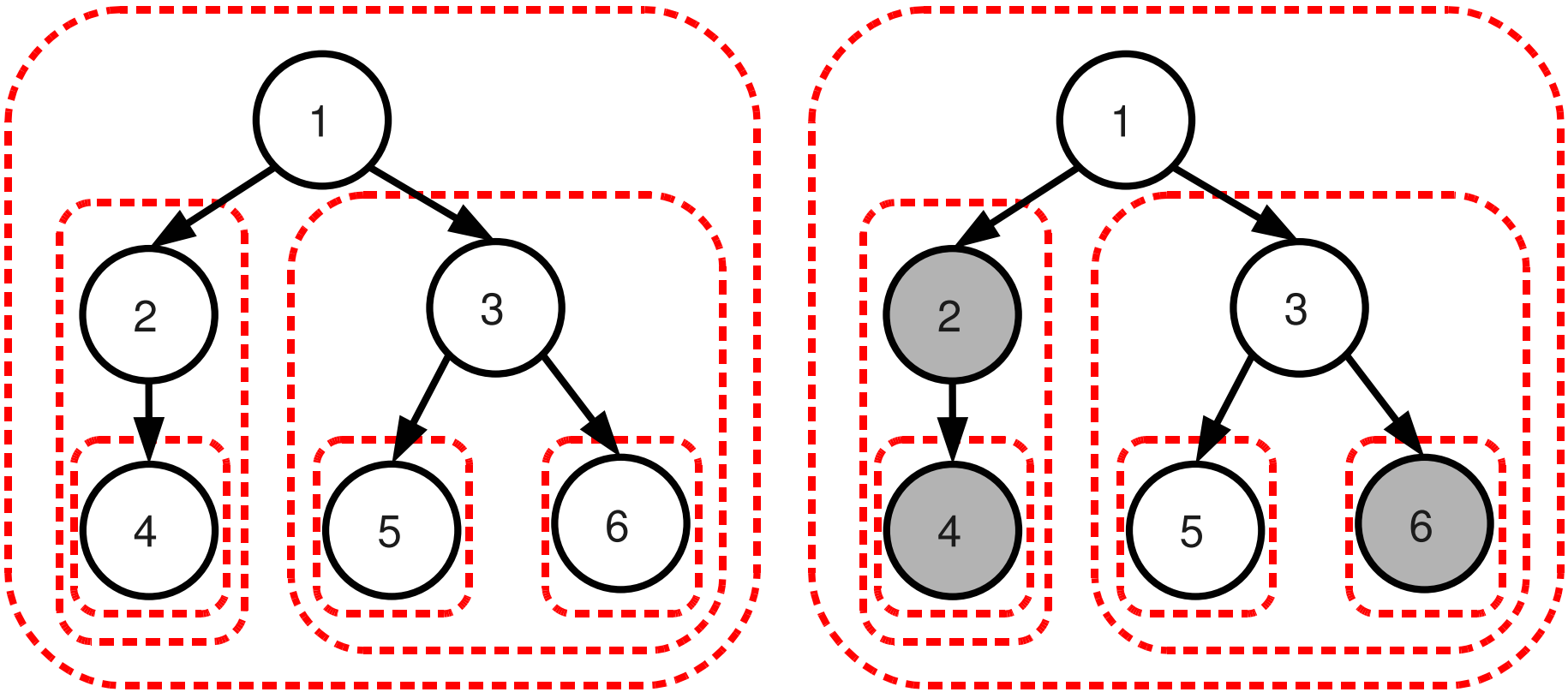
- the pattern of non-zero variables is an intersection of groups;
- the **zero pattern is a union of groups**.



$$\Omega(\beta) = \|\beta\|_2 + |\beta_2| + |\beta_3|.$$

Hierarchical Norms

[Zhao, Rocha, and Yu, 2009]



A node can be active only if its **ancestors are active**.

The selected patterns are **rooted subtrees.**

Proximal Gradient methods

A few proximal operators:

- ℓ_0-penalty: hard-thresholding;
- ℓ_1-norm: soft-thresholding;
- group-Lasso: group soft-thresholding;
- fused-lasso (1D total variation): [Hoefling, 2010];
- **hierarchical norms**: [Jenatton et al., 2010], $O(p)$ complexity;
- **overlapping group Lasso with ℓ_∞-norm**: [Mairal et al., 2010], (link with network flow optimization);

Modelling Patterns as Unions of Groups

the non-convex penalty of Huang, Zhang, and Metaxas [2009]

Warning: different point of view than in the two previous slides

$$\varphi(\boldsymbol{\beta}) \triangleq \min_{\mathcal{J} \subseteq \mathcal{G}} \left\{ \sum_{g \in \mathcal{J}} \eta_g \quad \text{s.t.} \quad \text{Supp}(\boldsymbol{\beta}) \subseteq \bigcup_{g \in \mathcal{J}} g \right\}.$$

- the penalty is **non-convex**.
- is **NP-hard** to compute (set cover problem).
- The pattern of non-zeroes in $\boldsymbol{\beta}$ is a **union** of (a few) groups.

It can be rewritten as a boolean linear program:

$$\varphi(\boldsymbol{\beta}) = \min_{\mathbf{x} \in \{0,1\}^{|\mathcal{G}|}} \left\{ \boldsymbol{\eta}^\top \mathbf{x} \quad \text{s.t.} \quad \mathbf{N}\mathbf{x} \geq \text{Supp}(\boldsymbol{\beta}) \right\}.$$

Modelling Patterns as Unions of Groups

convex relaxation and the penalty of Jacob, Obozinski, and Vert [2009]

The penalty of Huang et al. [2009]:

$$\varphi(\boldsymbol{\beta}) = \min_{\mathbf{x}\in\{0,1\}^{|\mathcal{G}|}} \left\{ \boldsymbol{\eta}^\top \mathbf{x} \;\text{ s.t. }\; \mathbf{N}\mathbf{x} \geq \text{Supp}(\boldsymbol{\beta}) \right\}.$$

A convex LP-relaxation:

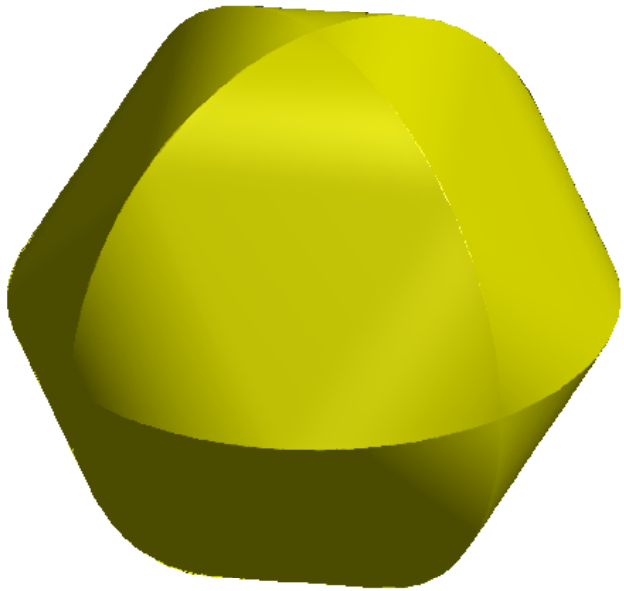
$$\psi(\boldsymbol{\beta}) \triangleq \min_{\mathbf{x}\in\mathbb{R}_+^{|\mathcal{G}|}} \left\{ \boldsymbol{\eta}^\top \mathbf{x} \;\text{ s.t. }\; \mathbf{N}\mathbf{x} \geq |\boldsymbol{\beta}| \right\}.$$

Lemma: ψ is the penalty of Jacob et al. [2009] with the ℓ_∞-norm:

$$\psi(\boldsymbol{\beta}) = \min_{(\boldsymbol{\xi}^g\in\mathbb{R}^p)_{g\in\mathcal{G}}} \sum_{g\in\mathcal{G}} \eta_g \|\boldsymbol{\xi}^g\|_\infty \;\text{ s.t. }\; \boldsymbol{\beta} = \sum_{g\in\mathcal{G}} \boldsymbol{\xi}^g \text{ and } \forall g,\ \text{Supp}(\boldsymbol{\xi}^g) \subseteq g,$$

Modelling Patterns as Unions of Groups

The norm of Jacob et al. [2009] in 3D



$\psi(\boldsymbol{\beta})$ with $\mathcal{G} = \{\{1,2\},\{2,3\},\{1,3\}\}$.

References I

M. Aharon, M. Elad, and A. M. Bruckstein. The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representations. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, November 2006.

R. G. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 2010. to appear.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

D. P. Bertsekas. *Nonlinear programming*. Athena Scientific Belmont, Mass, 1999.

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

References II

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

P. Bühlmann and B. Yu. Boosting with the L2 loss: regression and classification. *Journal of the American Statistical Association*, 98 (462):324–339, 2003.

T.T. Cai. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Annals of Statistics*, 27(3):898–924, 1999.

V. Cehver, M. F. Duarte, C. Hegde, and R. G. Baraniuk. Sparse signal recovery usingmarkov random fields. In *Advances in Neural Information Processing Systems*, 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.

F. Couzinie-Devy, J. Mairal, F. Bach, and J. Ponce. Dictionary learning for deblurring and digital zoom. *preprint arXiv:1110.0957*, 2011.

References III

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math*, 57:1413–1457, 2004.

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

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 54(12):3736–3745, December 2006.

K. Engan, S. O. Aase, and J. H. Husoy. Frame based signal compression using method of optimal directions (MOD). In *Proceedings of the 1999 IEEE International Symposium on Circuits Systems*, volume 4, 1999.

A. Haar. Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, 69:331–371, 1910.

References IV

H. Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010.

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 overlap 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. Technical report, 2009. preprint arXiv:0904.3523v1.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

References V

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, 2001.

M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000.

J. Mairal and B. Yu. Complexity analysis of the lasso regularization path. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69, January 2008a.

J. Mairal, G. Sapiro, and M. Elad. Learning multiscale sparse representations for image and video restoration. *SIAM Multiscale Modelling and Simulation*, 7(1):214–241, April 2008b.

References VI

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *ArXiv:0908.0050v1*, 2009. submitted.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network flow algorithms for structured sparsity. In *Advances in Neural Information Processing Systems*, 2010.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Convex and network flow optimization for structured sparsity. *Journal of Machine Learning Research*, 12:2681–2720, 2011.

S. Mallat. *A Wavelet Tour of Signal Processing, Second Edition*. Academic Press, New York, September 1999.

S. Mallat and Z. Zhang. Matching pursuit in a time-frequency dictionary. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

C.A. Micchelli, J.M. Morales, and M. Pontil. Regularizers for structured sparsity. *preprint arXiv:1010.0556v2*, 2011.

References VII

Y. Nesterov. A method for solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Math. Dokl.*, 27:372–376, 1983.

Y. Nesterov. *Introductory lectures on convex optimization*. Kluwer Academic Publishers, 2004.

Y. Nesterov. Gradient methods for minimizing composite objective function. Technical report, CORE, 2007.

J. Nocedal and SJ Wright. *Numerical Optimization*. Springer: New York, 2006. 2nd Edition.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37: 3311–3325, 1997.

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.

References VIII

Y.C. Pati, R. Rezaiifar, and PS Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, 1993.

M. Protter and M. Elad. Image sequence denoising via sparse and redundant representations. *IEEE Transactions on Image Processing*, 18(1):27–36, 2009.

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

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

References IX

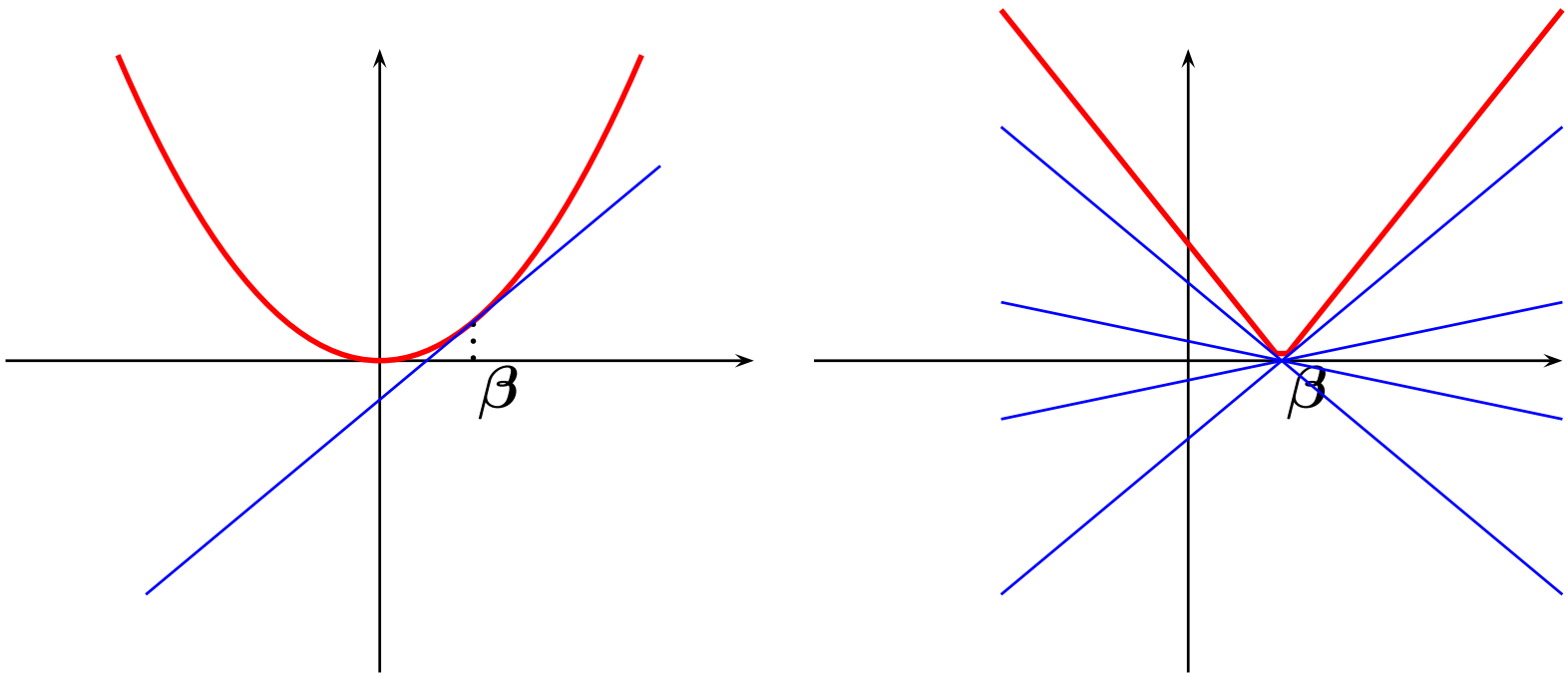
B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005.

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. 37(6A):3468–3497, 2009.

Appendix

Basic convex optimization tools: subgradients



(a) Smooth case (b) Non-smooth case

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

$$\partial f(\beta) \triangleq \{\kappa \in \mathbb{R}^p \mid f(\beta) + \kappa^\top(\beta' - \beta) \leq f(\beta') \text{ for all } \beta' \in \mathbb{R}^p\}.$$

Basic convex optimization tools: subgradients

Some nice properties

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

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: $\beta^\star$ minimizes g iff $\nabla g(\beta^\star) = 0$.
- g nondifferentiable: $\beta^\star$ minimizes g iff $0 \in \partial g(\beta^\star)$.

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 $\mathbb{R}^p$,

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

Exercises

- $\|\boldsymbol{\beta}\|_{**} = \|\boldsymbol{\beta}\|$ (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.
- $\partial\|\boldsymbol{\beta}\| = \{\boldsymbol{\kappa} \in \mathbb{R}^p \text{ s.t. } \|\boldsymbol{\kappa}\|_* \leq 1 \text{ and } \boldsymbol{\kappa}^\top\boldsymbol{\beta} = \|\boldsymbol{\beta}\|\}$.

Optimality conditions

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

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

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

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

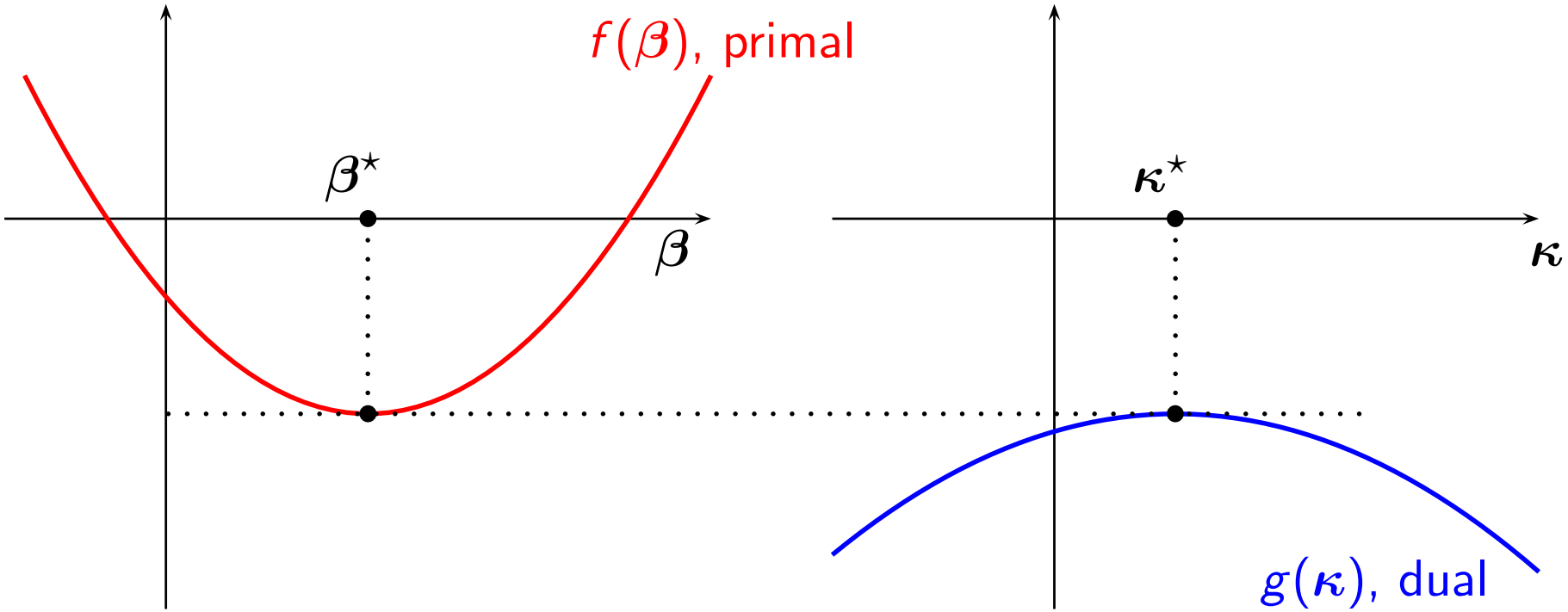
Since $\partial \|\boldsymbol{\beta}\| = \{\boldsymbol{\kappa} \in \mathbb{R}^p \text{ s.t. } \|\boldsymbol{\kappa}\|_* \leq 1 \text{ and } \boldsymbol{\kappa}^\top \boldsymbol{\beta} = \|\boldsymbol{\beta}\|\}$,

General optimality conditions:

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

Convex Duality

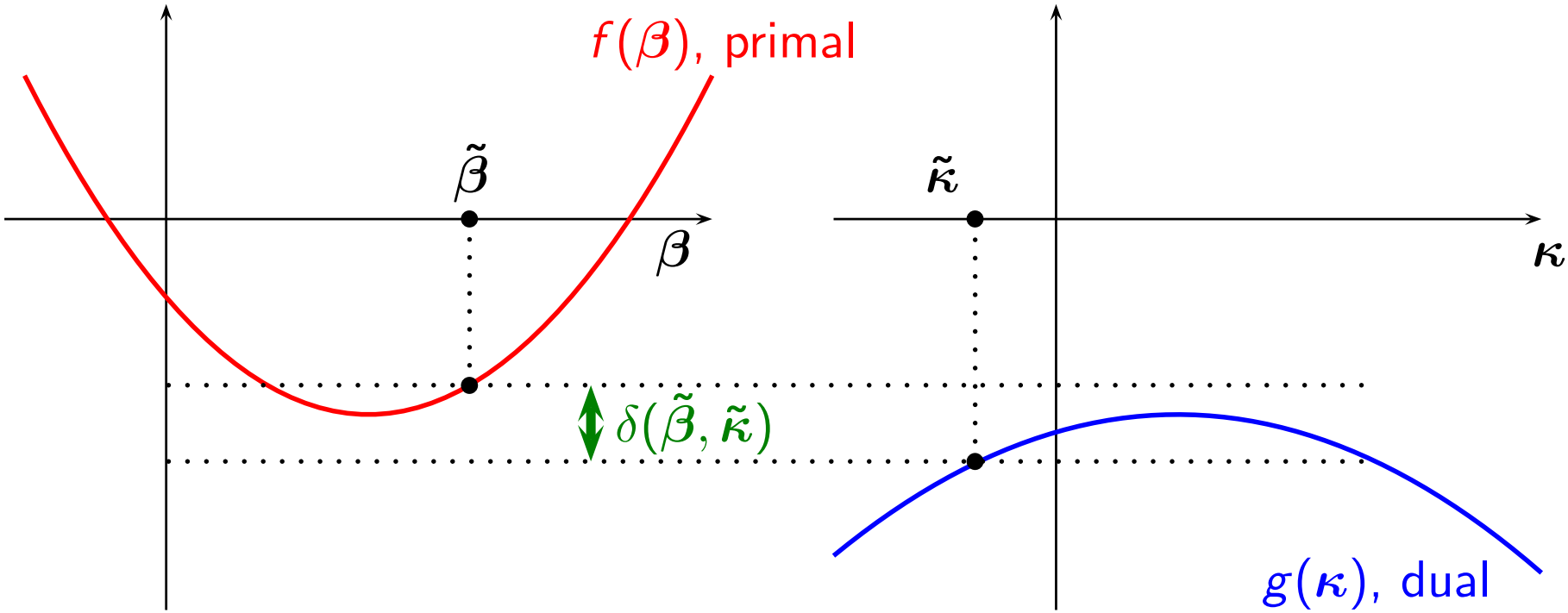
Strong Duality



Strong duality means that $\max_\kappa g(\kappa) = \min_\beta f(\beta)$

Convex Duality

Duality Gaps



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

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

Part V (Bonus): Dictionary Learning and Matrix Factorization

Matrix Factorization and Clustering

Let us cluster some training vectors $\mathbf{y}^1, \ldots, \mathbf{y}^m$ into p clusters using K-means:

$$\min_{(\mathbf{x}^j)_{j=1}^p, (l_i)_{i=1}^m} \sum_{i=1}^m \|\mathbf{y}^i - \mathbf{x}^{l_i}\|_2^2.$$

It can be equivalently formulated as

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times p}, \mathbf{B}\in\{0,1\}^{p\times m}} \sum_{i=1}^m \|\mathbf{y}^i - \mathbf{X}\boldsymbol{\beta}^i\|_F^2 \quad \text{s.t.} \quad \boldsymbol{\beta}^i \geq 0 \text{ and } \sum_{j=1}^p \boldsymbol{\beta}_j^i = 1,$$

which is a **matrix factorization** problem:

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times p}, \mathbf{B}\in\{0,1\}^{p\times m}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 \quad \text{s.t.} \quad \mathbf{B} \geq 0 \text{ and } \sum_{j=1}^p \boldsymbol{\beta}_j^i = 1,$$

Matrix Factorization and Clustering

Hard clustering

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times p},\mathbf{B}\in\{0,1\}^{p\times m}} \|\mathbf{Y}-\mathbf{X}\mathbf{B}\|_F^2 \quad \text{s.t.} \quad \mathbf{B}\geq 0 \text{ and } \sum_{j=1}^{p}\beta_j^i = 1,$$

$\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_p]$ are the centroids of the p clusters.

Soft clustering

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times p},\mathbf{B}\in\mathbb{R}^{p\times m}} \|\mathbf{Y}-\mathbf{X}\mathbf{B}\|_F^2 \quad \text{s.t.} \quad \mathbf{B}\geq 0 \text{ and } \sum_{j=1}^{p}\beta_j^i = 1,$$

$\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_p]$ are the centroids of the p clusters.

Other Matrix Factorization Problems

PCA

$$\min_{\substack{\mathbf{B}\in\mathbb{R}^{p\times n}\\ \mathbf{X}\in\mathbb{R}^{m\times p}}} \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\mathbf{B}\|_F^2 \quad \text{s.t.} \quad \mathbf{X}^\top\mathbf{X}=\mathbf{I} \text{ and } \mathbf{B}\mathbf{B}^\top \text{ is diagonal.}$$

$\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_p]$ are the principal components.

Other Matrix Factorization Problems

Non-negative matrix factorization [Lee and Seung, 2001]

$$\min_{\substack{\mathbf{B}\in\mathbb{R}^{p\times n}\\ \mathbf{X}\in\mathbb{R}^{m\times p}}} \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\mathbf{B}\|_F^2 \quad \text{s.t.} \quad \mathbf{B}\geq 0 \text{ and } \mathbf{X}\geq 0.$$

Dictionary Learning and Matrix Factorization

[Olshausen and Field, 1997]

$$\min_{\mathbf{X}\in\mathcal{X},\mathbf{B}\in\mathbb{R}^{p\times m}} \sum_{i=1}^{n} \frac{1}{2}\|\mathbf{y}^i - \mathbf{X}\boldsymbol{\beta}^i\|_F^2 + \lambda\|\boldsymbol{\beta}^i\|_1,$$

which is again a matrix factorization problem

$$\min_{\substack{\mathbf{B}\in\mathbb{R}^{p\times n}\\ \mathbf{X}\in\mathcal{X}}} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + \lambda\|\mathbf{B}\|_1.$$

Why having a unified point of view?

$$\min_{\substack{\mathbf{B}\in\mathcal{B}\\ \mathbf{X}\in\mathcal{X}}} \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\mathbf{B}\|_F^2 + \lambda\psi(\mathbf{B}).$$

- same framework for NMF, sparse PCA, dictionary learning, clustering, topic modelling;
- can play with various constraints/penalties on $\mathbf{B}$ (coefficients) and on $\mathbf{X}$ (loadings, dictionary, centroids);
- same algorithms (no need to reinvent the wheel): alternate minimization, online learning [Mairal et al., 2009].

The Image Denoising Problem

$$\underbrace{\mathbf{y}}_{\text{measurements}} = \underbrace{\mathbf{x}_{orig}}_{\text{original image}} + \underbrace{\varepsilon}_{\text{noise}}$$

Sparse representations for image restoration

$$\underbrace{\mathbf{y}}_{\text{measurements}} = \underbrace{\mathbf{x}_{orig}}_{\text{original image}} + \underbrace{\varepsilon}_{\text{noise}}$$

Energy minimization problem - MAP estimation

$$E(\mathbf{x}) = \underbrace{\frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2}_{\text{relation to measurements}} + \underbrace{\psi(\mathbf{x})}_{\text{image model (-log prior)}}$$

Some classical priors

- Smoothness $\lambda\|\mathcal{L}\mathbf{x}\|_2^2$
- Total variation $\lambda\|\nabla\mathbf{x}\|_1^2$
- MRF priors
- ...

Sparse representations for image restoration

Designed dictionaries

[Haar, 1910], [Zweig, Morlet, Grossman ∼70s], [Meyer, Mallat, Daubechies, Coifman, Donoho, Candes ∼80s-today]. . . (see [Mallat, 1999])
Wavelets, Curvelets, Wedgelets, Bandlets, . . . lets

Learned dictionaries of patches

[Olshausen and Field, 1997], [Engan et al., 1999], [Lewicki and Sejnowski, 2000], [Aharon et al., 2006]

$$\min_{\boldsymbol{\beta}_i, \mathbf{X} \in \mathcal{C}} \sum_i \underbrace{\frac{1}{2}\|\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i\|_2^2}_{\text{reconstruction}} + \underbrace{\lambda\psi(\boldsymbol{\beta}_i)}_{\text{sparsity}}$$

- $\psi(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_0$ ("ℓ_0 pseudo-norm")
- $\psi(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ (ℓ_1 norm)

Sparse representations for image restoration

Solving the denoising problem

[Elad and Aharon, 2006]

- Extract all overlapping 8×8 patches $\mathbf{y}_i$.
- Solve a matrix factorization problem:

$$\min_{\boldsymbol{\beta}_i, \mathbf{X} \in \mathcal{C}} \sum_{i=1}^{n} \underbrace{\frac{1}{2}\|\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i\|_2^2}_{\text{reconstruction}} + \underbrace{\lambda\psi(\boldsymbol{\beta}_i)}_{\text{sparsity}},$$

  with $n > 100,000$

- Average the reconstruction of each patch.

Sparse representations for image restoration

K-SVD: [Elad and Aharon, 2006]

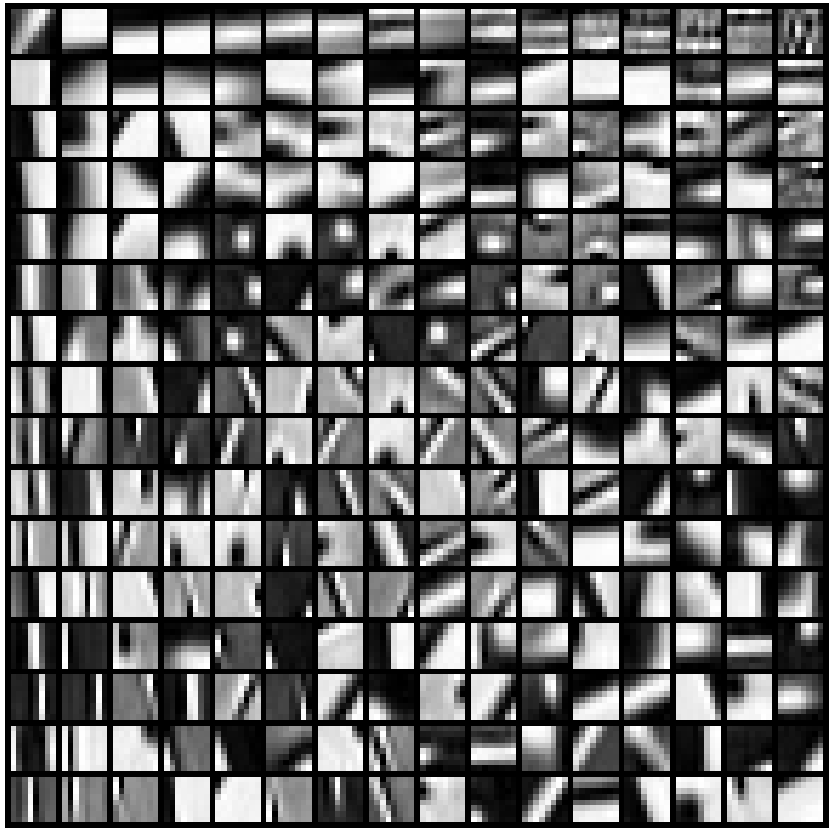
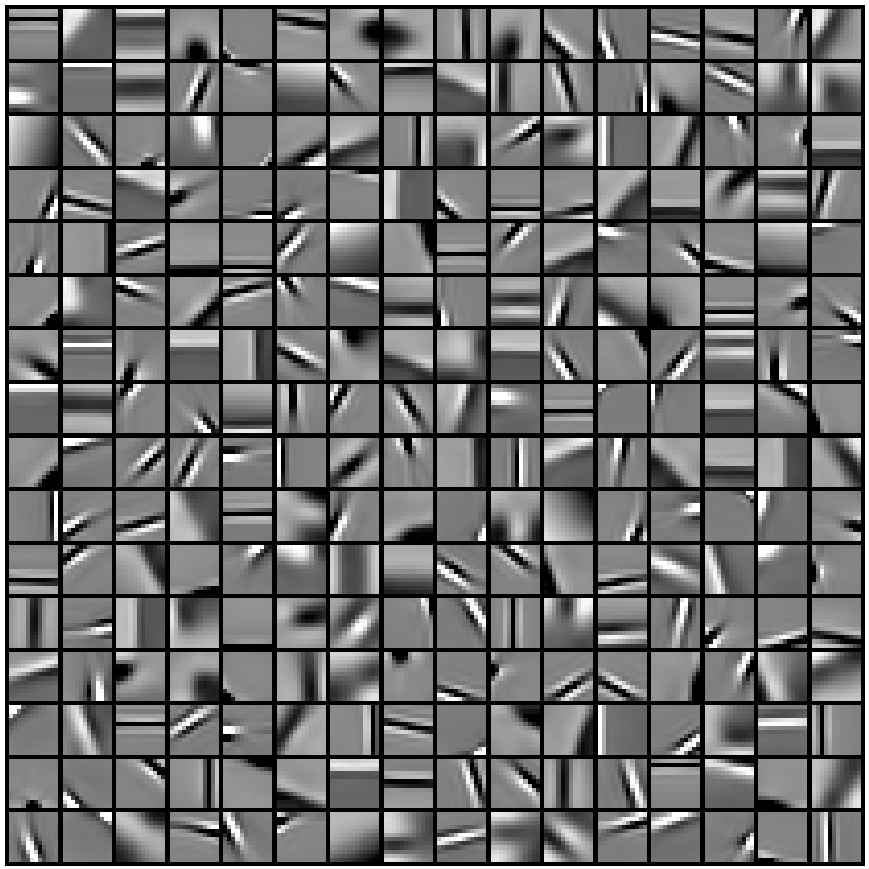
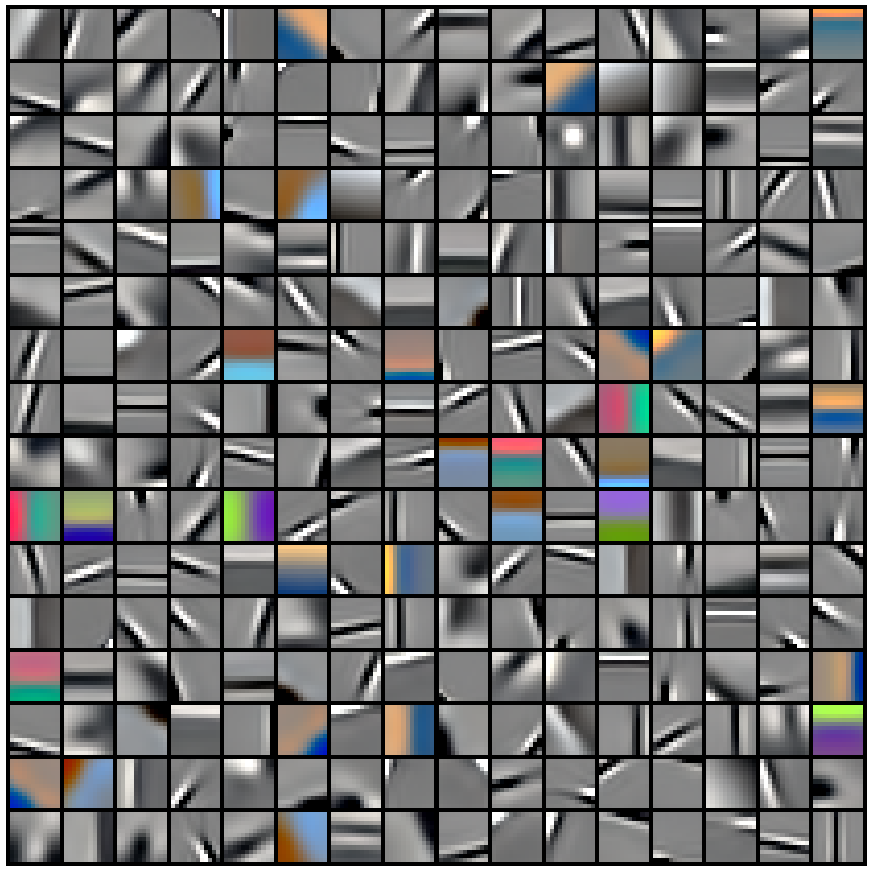


Figure: Dictionary trained on a noisy version of the image boat.

Sparse representations for image restoration

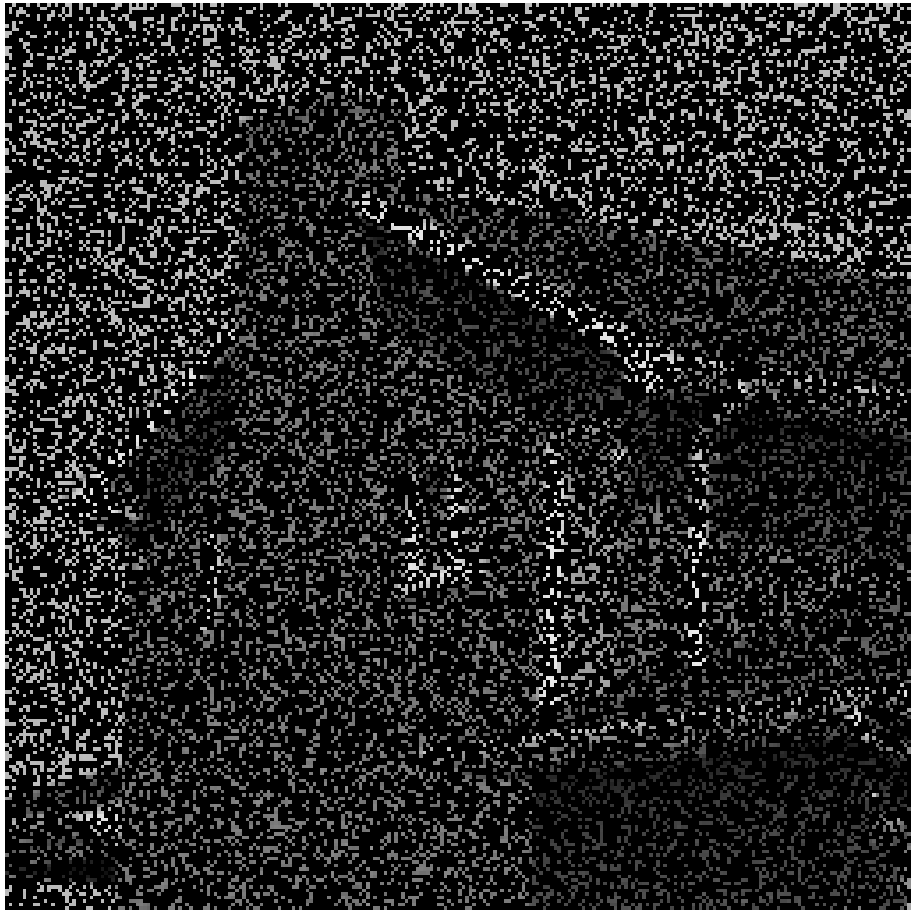
Grayscale vs color image patches





Sparse representations for image restoration

[Mairal, Sapiro, and Elad, 2008b]



Sparse representations for image restoration

Inpainting, [Mairal, Elad, and Sapiro, 2008a]

Sparse representations for image restoration

Inpainting, [Mairal, Elad, and Sapiro, 2008a]

Sparse representations for video restoration

Key ideas for video processing

[Protter and Elad, 2009]

- Using a 3D dictionary.
- Processing of many frames at the same time.
- Dictionary propagation.

Sparse representations for image restoration

Inpainting, [Mairal, Sapiro, and Elad, 2008b]

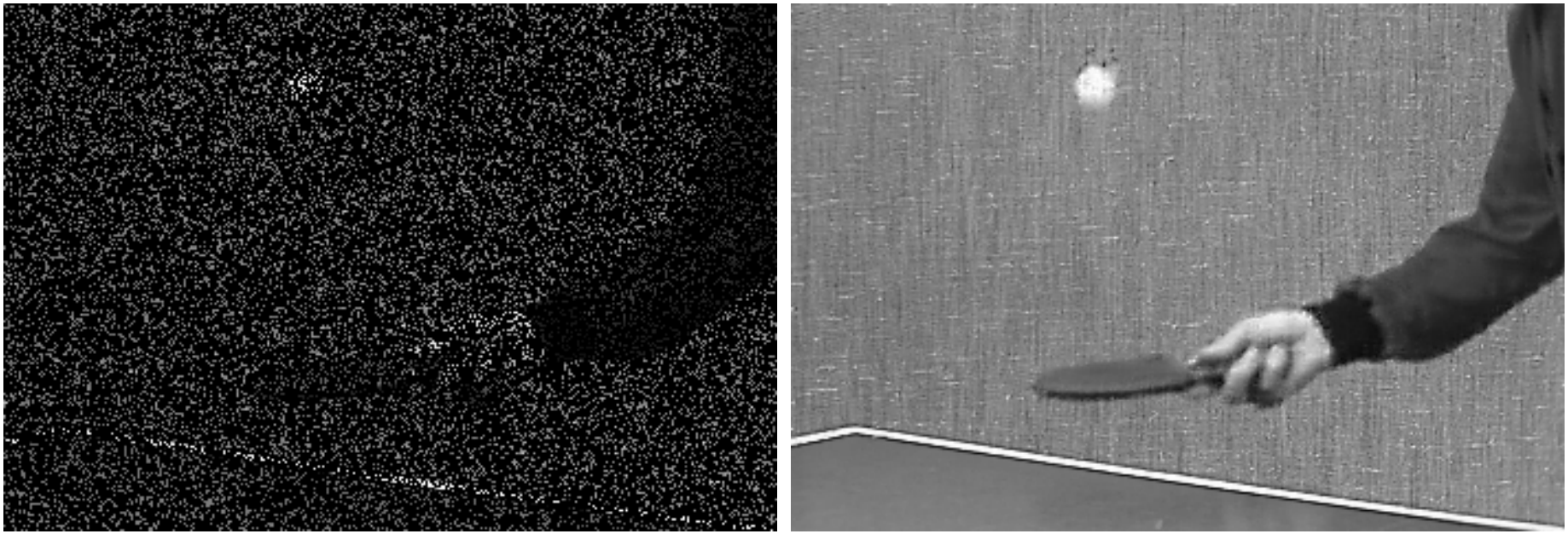


Figure: Inpainting results.

Sparse representations for image restoration

Inpainting, [Mairal, Sapiro, and Elad, 2008b]

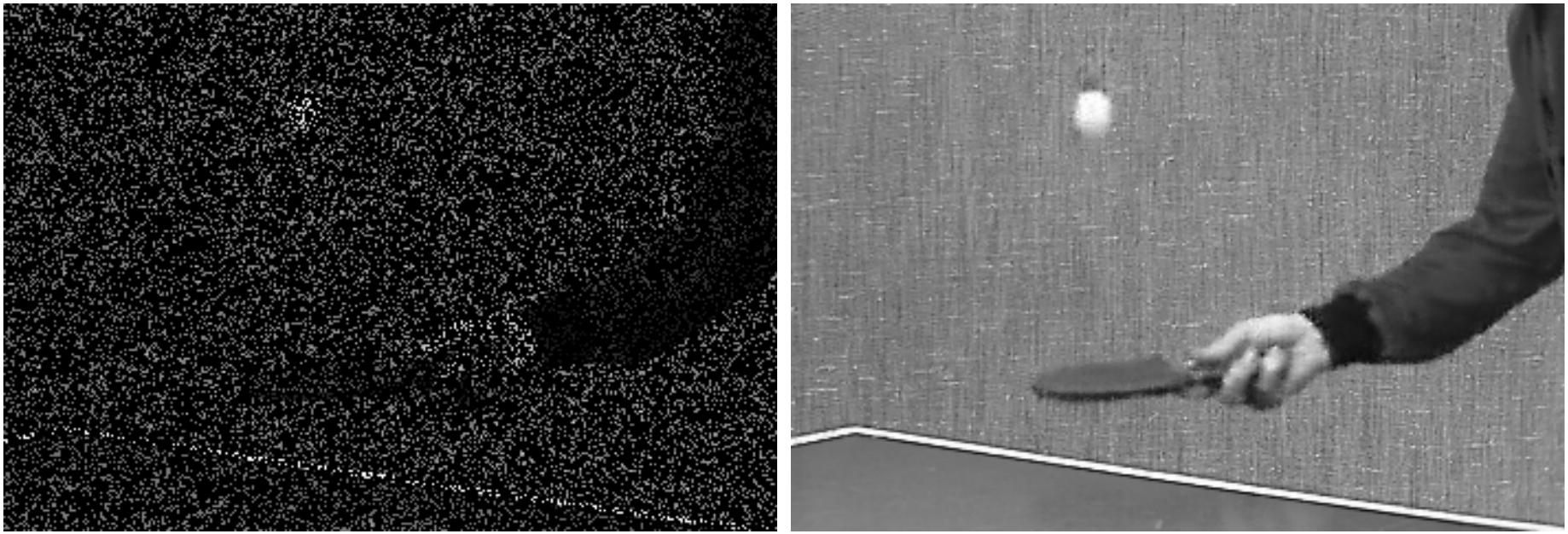


Figure: Inpainting results.

Sparse representations for image restoration

Inpainting, [Mairal, Sapiro, and Elad, 2008b]

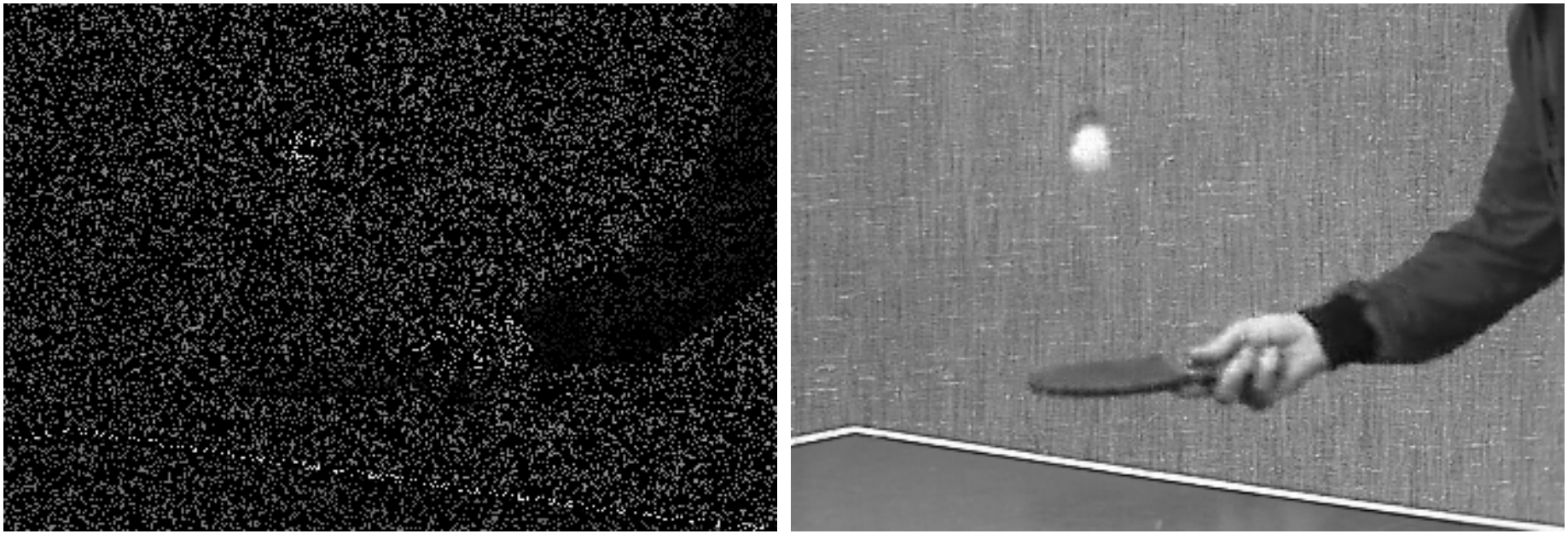


Figure: Inpainting results.

Sparse representations for image restoration

Inpainting, [Mairal, Sapiro, and Elad, 2008b]

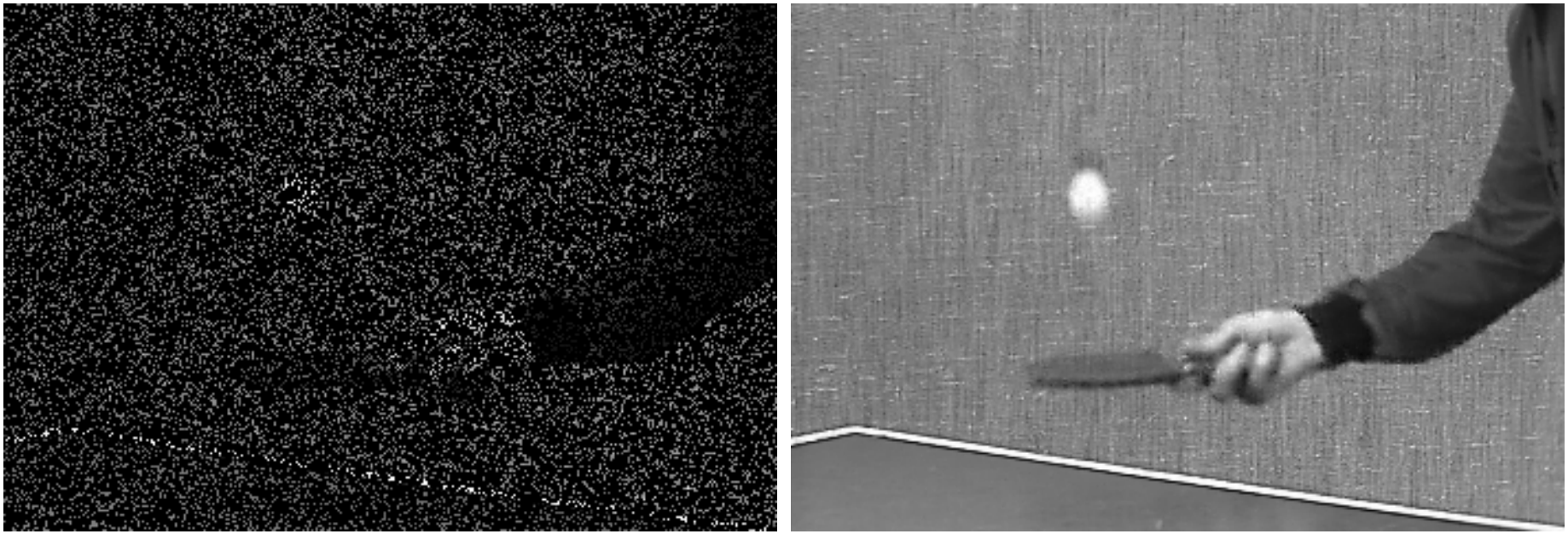


Figure: Inpainting results.

Sparse representations for image restoration

Inpainting, [Mairal, Sapiro, and Elad, 2008b]

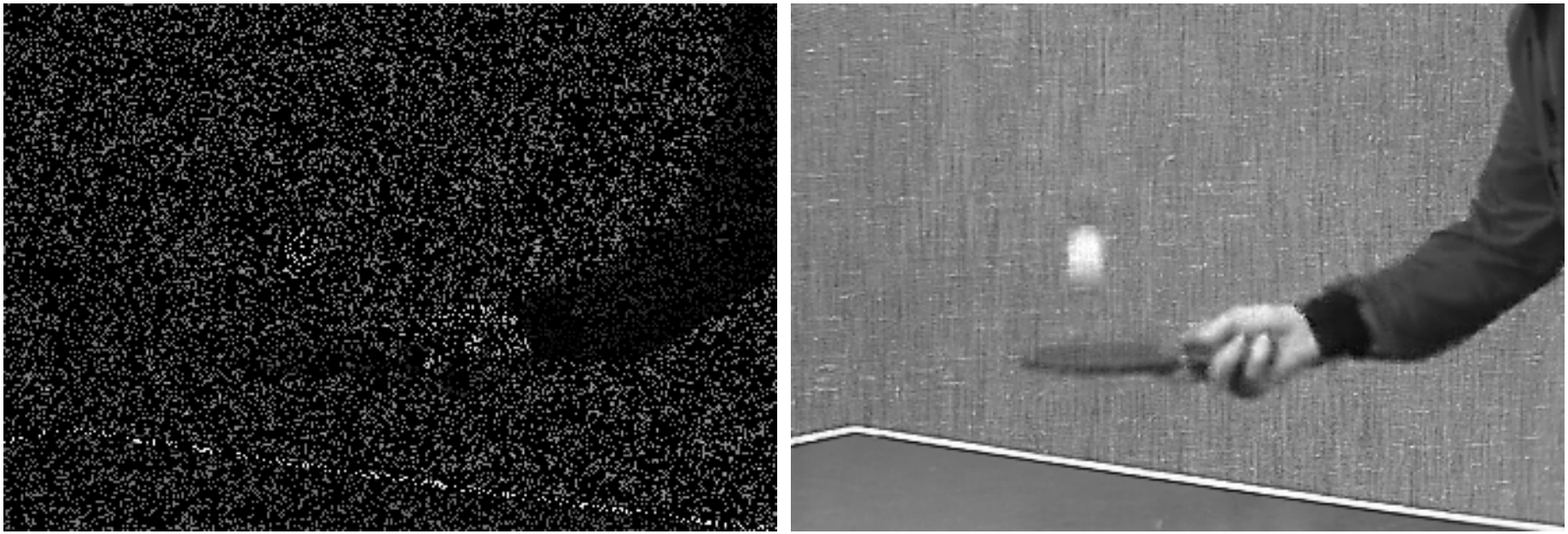


Figure: Inpainting results.

Sparse representations for image restoration

Color video denoising, [Mairal, Sapiro, and Elad, 2008b]

Figure: Denoising results. $\sigma = 25$

Sparse representations for image restoration

Color video denoising, [Mairal, Sapiro, and Elad, 2008b]

Figure: Denoising results. $\sigma = 25$

Sparse representations for image restoration

Color video denoising, [Mairal, Sapiro, and Elad, 2008b]

Figure: Denoising results. $\sigma = 25$

Sparse representations for image restoration

Color video denoising, [Mairal, Sapiro, and Elad, 2008b]

Figure: Denoising results. $\sigma = 25$

Sparse representations for image restoration

Color video denoising, [Mairal, Sapiro, and Elad, 2008b]

Figure: Denoising results. $\sigma = 25$

Digital Zooming

[Couzinie-Devy et al., 2011], Original

Digital Zooming

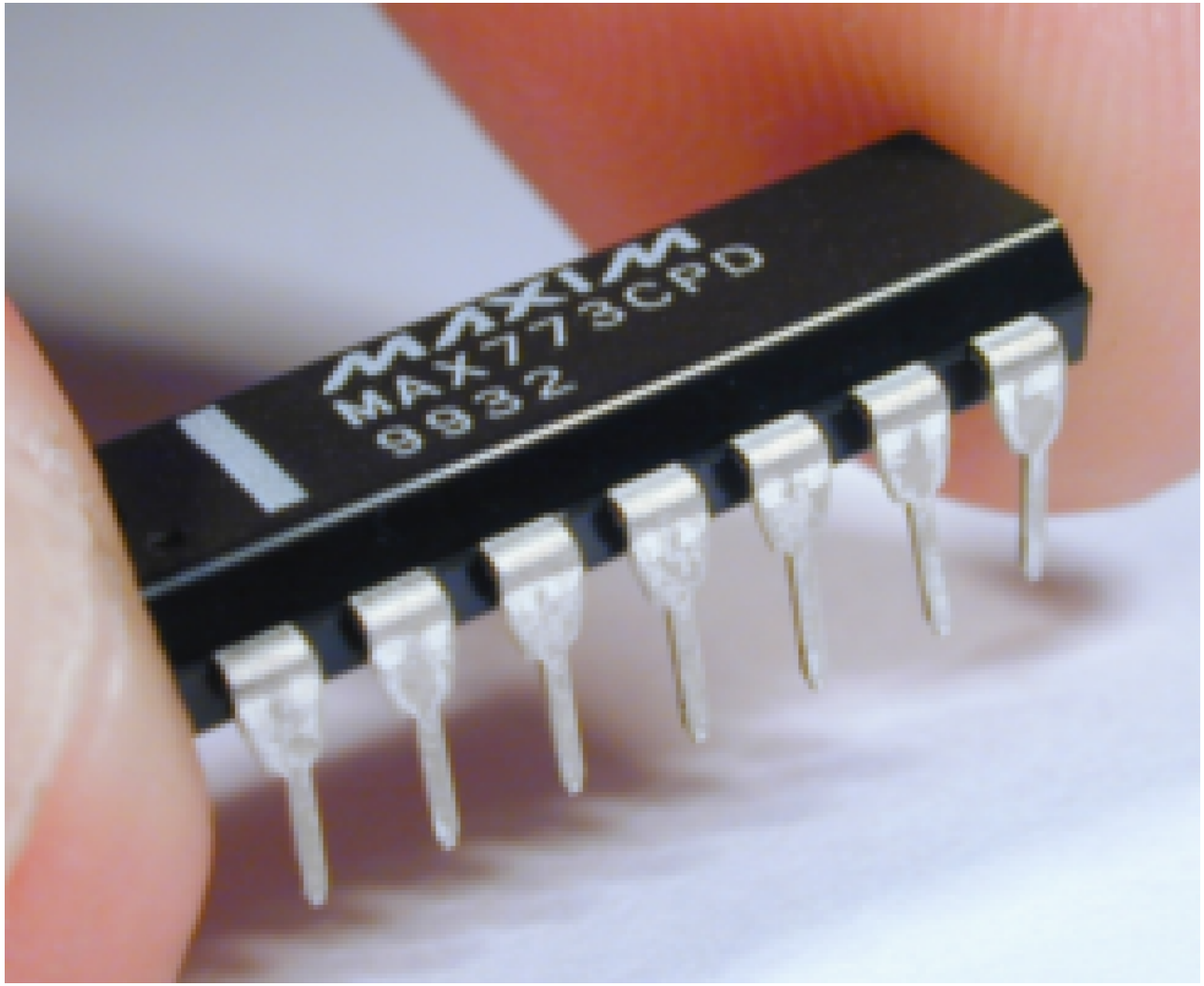
[Couzinie-Devy et al., 2011], Bicubic

Digital Zooming

[Couzinie-Devy et al., 2011], Proposed method

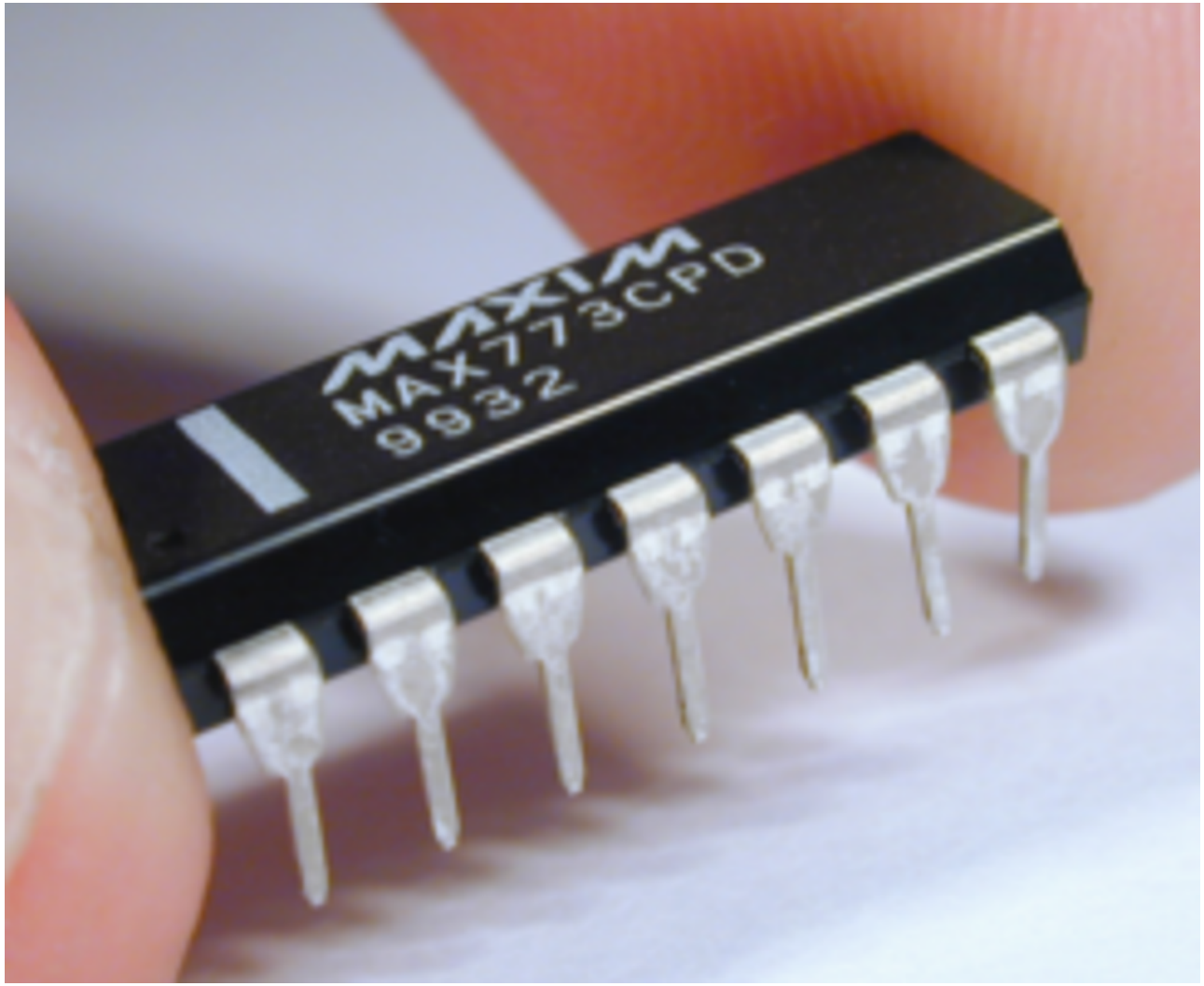
Digital Zooming

[Couzinie-Devy et al., 2011], Original



Digital Zooming

[Couzinie-Devy et al., 2011], Bicubic



Digital Zooming

[Couzinie-Devy et al., 2011], Proposed approach

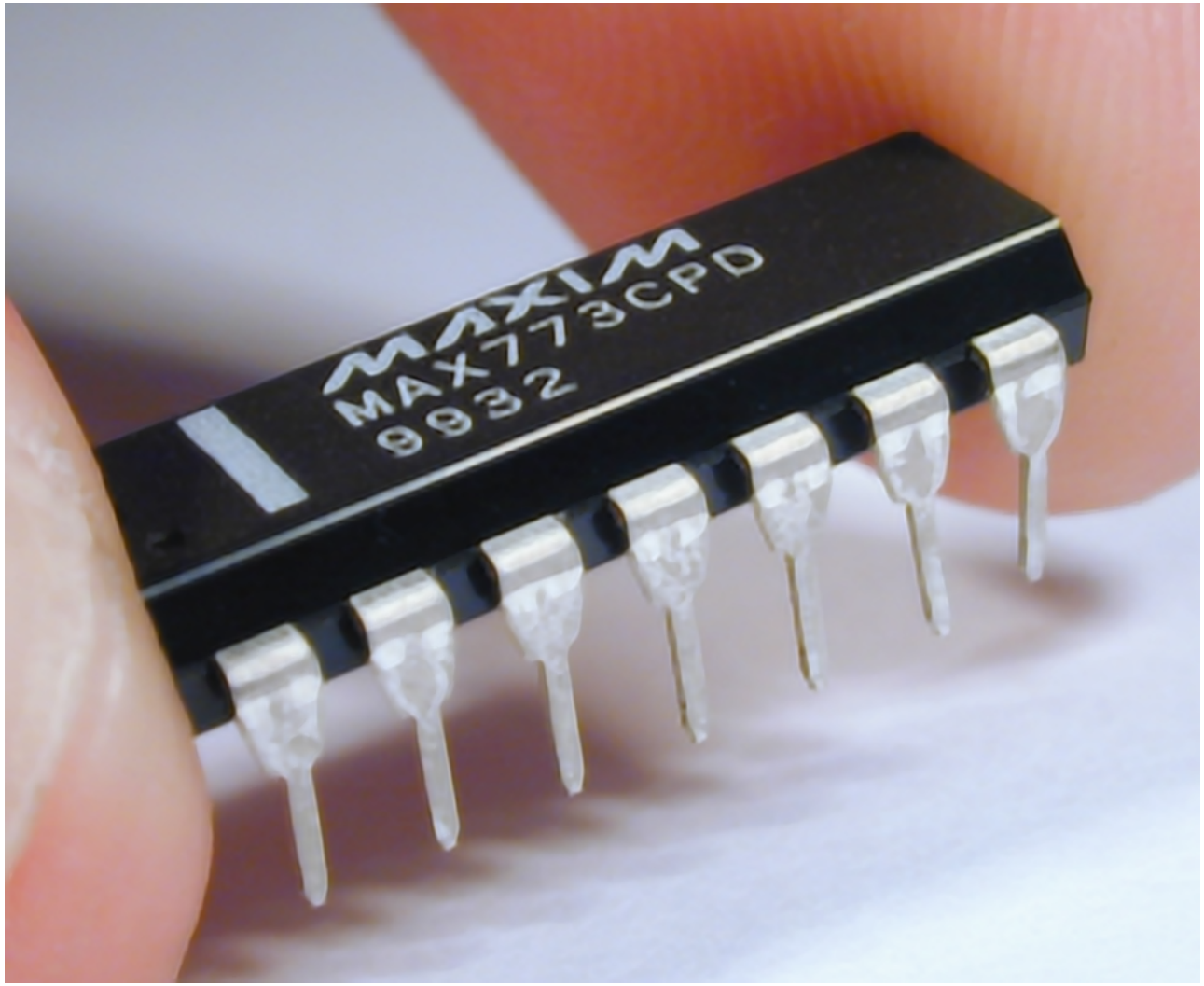


Image Deblurring

[Couzinie-Devy et al., 2011], Original

Image Deblurring

[Couzinie-Devy et al., 2011], Blurry and Noisy

Image Deblurring

[Couzinie-Devy et al., 2011], Result

Image Deblurring

[Couzinie-Devy et al., 2011], Original

Image Deblurring

[Couzinie-Devy et al., 2011], Blurry and Noisy

Image Deblurring

[Couzinie-Devy et al., 2011], Result

Inverse half-toning

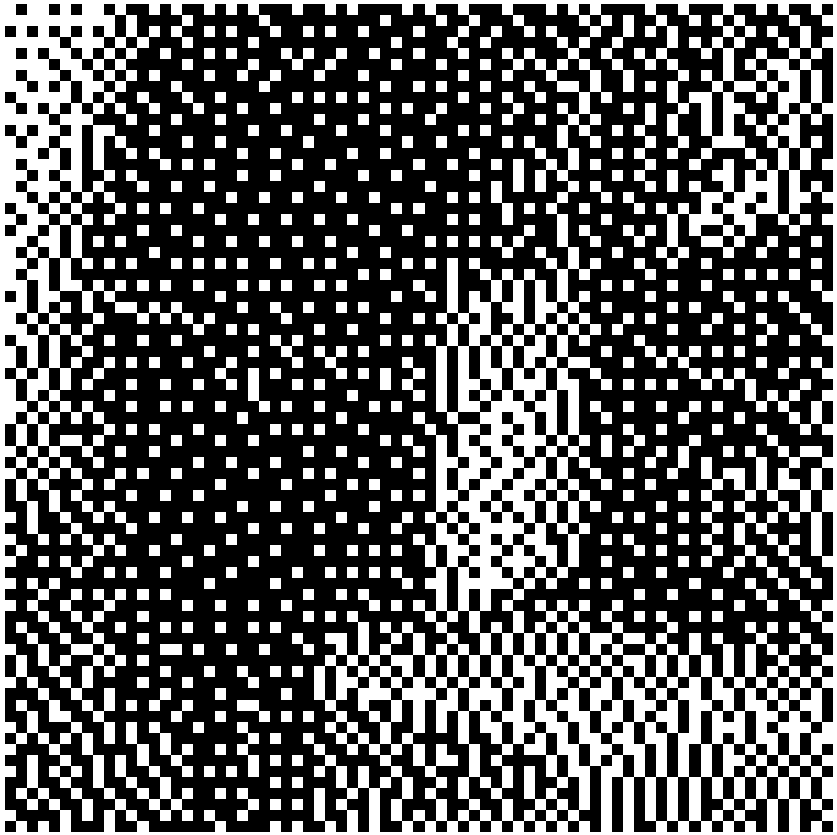
Original

Inverse half-toning

Reconstructed image

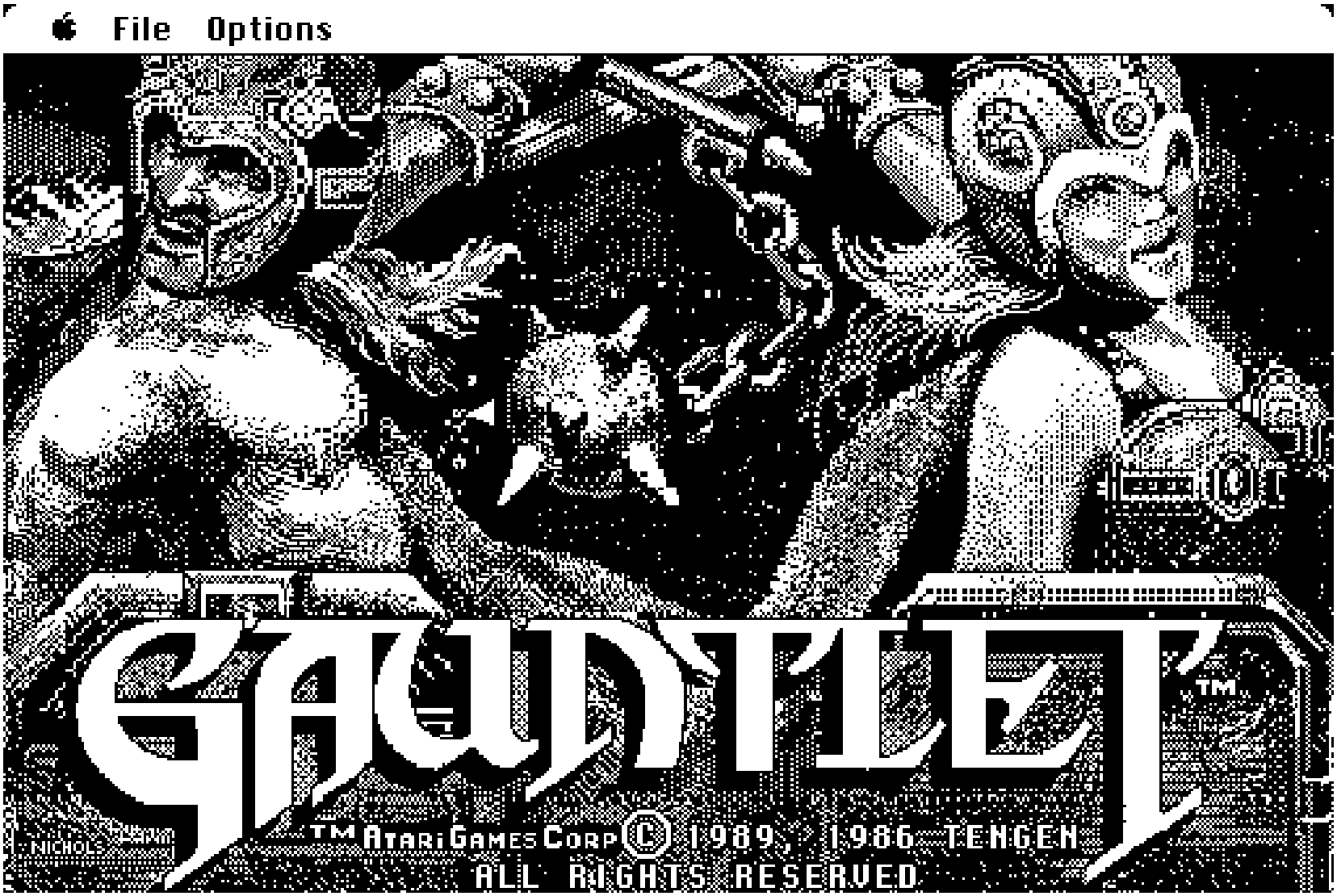
Inverse half-toning

Reconstructed image



Inverse half-toning

Original



Inverse half-toning

Reconstructed image

Inverse half-toning

Original

Inverse half-toning

Reconstructed image

Inverse half-toning

Original

Inverse half-toning

Reconstructed image

Inverse half-toning

Original

Inverse half-toning

Reconstructed image