# $\ell^1$-regularization

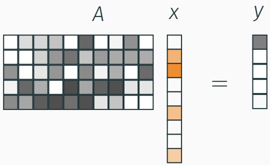
Fallstudien der mathematischen Modellbildung, Teil 2

20.10.2023 - 21.11.2023,

paul.catala@tum.de

## Sparsity

- Enforcing structure helps with ill-conditioning and under-determined systems.
  A popular prior is sparsity, *i.e.* assuming the solution has only a few non-zero entries




- **Rationale**: signals/data are often sparse in some basis / living on low-complexity domain.

- If $c_i$, $i = 1, \ldots, n$ denotes the columns of $A$, the system rewrites

$$y = \sum_{i=1}^{n} x_i c_i$$

$(c_i)$ is an over-complete basis (or dictionary), and the goal is to select a subset of this basis that is sufficient to express $y \to$ regressor selection, or variable selection.

- If $c_i$, $i = 1, \dots, n$ denotes the columns of $A$, the system rewrites

$$y = \sum_{i=1}^{n} x_i c_i$$

  $(c_i)$ is an over-complete basis (or dictionary), and the goal is to select a subset of this basis that is sufficient to express $y \to$ regressor selection, or variable selection.

- A natural candidate to promote sparsity of solutions is the $\ell^0$-norm

$$\|x\|_0 = \# \{i \in \{1, \dots, n\} \; ; \; x_i \neq 0\}$$

  ! It is actually not a norm !

- If $c_i$, $i = 1, \ldots, n$ denotes the columns of $A$, the system rewrites

$$y = \sum_{i=1}^{n} x_i c_i$$

  $(c_i)$ is an over-complete basis (or dictionary), and the goal is to select a subset of this basis that is sufficient to express $y \to$ regressor selection, or variable selection.

- A natural candidate to promote sparsity of solutions is the $\ell^0$-norm

$$\|x\|_0 = \# \{i \in \{1, \ldots, n\} \; ; \; x_i \neq 0\}$$

  ! It is actually not a norm !

- The corresponding regularized problem is

$$\boxed{\min \|Ax - y\|^2 + \lambda \|x\|_0}$$

  and in the noiseless case

$$\boxed{\min \|x\|_0 \quad \text{s.t.} \quad Ax = y}$$

- Remember that the penalized form is always equivalent to a constrained form with adequate parameter, *i.e.*

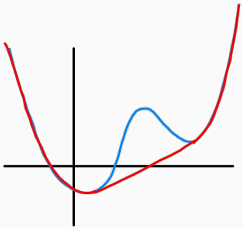
$$\min_{x \in \mathbb{R}^n} \|Ax - y\|^2 \quad \text{s.t.} \quad \|x\|_0 \leqslant \tau \tag{1}$$

- NP-hard combinatorial, non-convex problem. Direct strategy: check every possible sparsity pattern, *i.e.* fix subsets $J$ of non-zero entries in $x$ and solve the least-squares

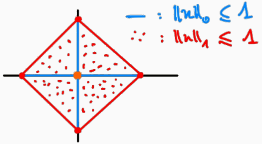
$$\min_{\tilde{x} \in \mathbb{R}^n} \|A_J \tilde{x} - y_J\|^2$$

  There are $\binom{n}{k}$ possible supports for each sparsity level $\rightarrow$ infeasible for large $n$
- Possible approximations of the problem:
  - Greedy algorithms (*e.g.* orthogonal matching pursuit)
  - Convex relaxation

- *Definition (Convex envelope).* The convex envelope of a function $I(x)$ is the largest convex $J(x)$ such that $J(x) \leqslant I(x)$ .



- ***Theorem.*** The convex envelope of $\|x\|_0$ for $x$ restricted to $\|x\|_\infty \leqslant \alpha$ is $\|x\|_1/\alpha$



# Convex Relaxation

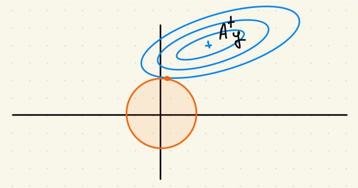
- Relax $\ell^0$-penalty into $\ell^1$-penalty

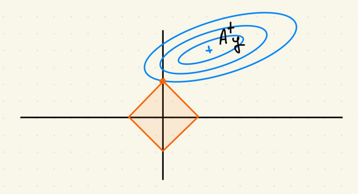
$$\min \|Ax - y\|_2^2 + \lambda \|x\|_1 \qquad \text{(Lasso)}$$

Called Lasso[1] (Least Absolute Shrinkage and Selection Operator) or basis pursuit denoising. When $\lambda = 0$, we obtain the basis pursuit[2] problem

$$\min \|x\|_1 \quad \text{s.t.} \quad Ax = y \qquad \text{(BP)}$$

- Main properties are:
Shrinkage: like Tikhonov regularization, Lasso penalizes large coefficients
Selection: unlike Tikhonov, Lasso produces sparse estimates




[1] *Tibshirani, 1996*

[2] *Donoho, early 1990's*

- We can reformulate the problem under a constrained form

$$\min \frac{1}{2}\|z-y\|^2 + \lambda\|x\|_1 \quad \text{s.t.} \quad z = Ax$$

and deduce the Lagrangian:

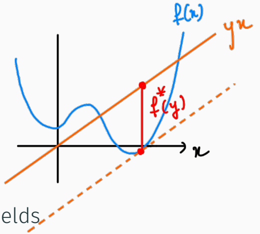
$$\mathcal{L}(x,z,\nu) = \frac{1}{2}\|z-y\|^2 + \lambda\|x\|_1 + \nu^\top(z - Ax)$$

- Minimization over $z$ yields $\tilde{z} = y - \nu$. Minimization over $x$ on the other hand is less obvious, since we have lost differentiability

$$\inf_x \lambda\|x\|_1 - \langle A^\top \nu, x\rangle = -\left(\sup_x \langle A^\top\nu, x\rangle - \lambda\|x\|_1\right)$$

***Definition (Conjugate function).*** The convex conjugate of $f : \mathbb{R}^n \to \mathbb{R}$ is

$$f^*(y) = \sup_x \langle y, x\rangle - f(x)$$




With $J(x) := \lambda\|x\|_1$, the minimization over $x$ and $z$ yields

$$\mathcal{L}(\tilde{x},\tilde{z},\nu) = \nu^\top y - \frac{1}{2}\|\nu\|^2 - J^*(A^\top\nu)$$

- ***Definition (Dual norm).*** Given a norm $\|\cdot\|$ on $\mathbb{R}^n$, the associated dual norm is

$$\|y\|_* = \sup\left\{y^\top x \; ; \; \|x\| \leqslant 1\right\}$$

*Example.* $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are dual to each other.

- ***Proposition.*** The conjugate function of $\|x\|$ is

$$f^*(y) = \begin{cases} 0 & \text{if} \quad \|y\|_* \leqslant 1 \\ \infty & \text{otherwise} \end{cases}$$

**Proof.**[1] If $\|y\|_* > 1$, then by definition there exists $w \in \mathbb{R}^n$ such that $\|w\| \leqslant 1$ and $y^\top w > 1$. Taking $x = tw$ and letting $t \to \infty$ we obtain

$$y^\top x - \|x\| = t(y^\top w - \|w\|) \to \infty,$$

hence $f^*(y) = \infty$. If $\|y\|_* \leqslant 1$, since $y^\top x \leqslant \|x\|\|y\|_*$ for all $x$, then $y^\top x - \|x\| \leqslant 0$, and $x = 0$ is the maximizer.

[1] *Boyd, Vandenberghe, Convex Optimization, Example 3.26*

# Dual Lasso

- If $J(x) = \lambda\|x\|_1$, then $J^*(y)$ is the indicator of $\{\|y\|_\infty \leqslant \lambda\}$.
- Altogether, we obtain

$$\mathcal{L}(\tilde{x}, \tilde{z}, \nu) = \nu^\top y - \frac{1}{2}\|\nu\|^2 - i_{\{\nu : \|\nu\|_\infty \leqslant \lambda\}}(A^\top \nu)$$

where we denote $i_C$ the indicator function of the set $C$. Hence the Lasso dual problem reads

$$\boxed{\max \nu^\top y - \frac{1}{2}\|\nu\|^2 \quad \text{s.t.} \quad \|A^\top \nu\|_\infty \leqslant \lambda}$$

# Subdifferential

$\|\cdot\|_1$ is convex but not differentiable at 0. How to derive optimality conditions?

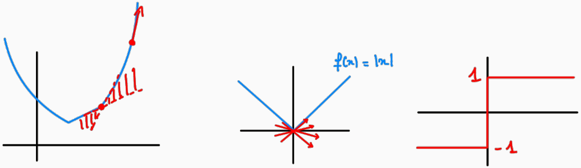
- Recall the standard inequality for convex functions

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle$$

- ***Definition (Sub-differential).*** The sub-differential of $f : \mathbb{R}^n \to \mathbb{R}$ at $x$ is

$$\partial f(x) = \left\{ v \in \mathbb{R}^n \ ; \ \forall y \in \mathbb{R}^n, f(y) \geqslant f(x) + \langle v, y - x \rangle \right\}$$

Note that $\partial f(x)$ is convex. If $f$ is differentiable, then $\partial f(x) = \{\nabla f(x)\}$.




- ***Proposition.*** For *any* function $f$,

$$x_* = \operatorname{argmin}_x f(x) \iff 0 \in \partial f(x)$$

**Proof.** $x_*$ minimizer of $f$ $\iff$ $\forall x, f(x) \geqslant f(x_*) = f(x_*) + \langle 0, x - x^* \rangle$ $\iff$ $0 \in \partial f(x)$.

Some basic rules

- $\partial f(x) = \{\nabla f(x)\}$ if $f$ is differentiable at $x$
- $\partial(\alpha f) = \alpha \partial f$ if $\alpha > 0$
- $\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x)$ (except in pathological cases: according to Moreau-Rockafellar theorem, if there exists a point $x_0 \in \mathrm{dom}(f_1 + f_2)$ such that $f_1$ is continuous at $x_0$, then the equality holds for any $x \in \mathrm{dom}(f_1 + f_2)$).
- if $g(x) = f(Ax + b)$ where $f$ is convex, then $\partial g(x) = A^\top \partial f(Ax + b)$

# Subdifferential of $\|\cdot\|_1$

- $|x|$ is differentiable at any $x \neq 0$ with derivative $\pm 1$. At 0,

$$(\forall z \in \mathbb{R}, |z| \geqslant yz) \iff y \in [-1, 1]$$

so $\partial 0 = [-1, 1]$, and

$$\partial |x| = \begin{cases} \{1\} & \text{if} \quad x > 0 \\ [-1, 1] & \text{if} \quad x = 0 \\ \{-1\} & \text{if} \quad x < 0 \end{cases}$$

- $|x|$ is differentiable at any $x \neq 0$ with derivative $\pm 1$. At 0,

$$(\forall z \in \mathbb{R}, |z| \geqslant yz) \iff y \in [-1, 1]$$

so $\partial 0 = [-1, 1]$, and

$$\partial |x| = \begin{cases} \{1\} & \text{if} \quad x > 0 \\ [-1, 1] & \text{if} \quad x = 0 \\ \{-1\} & \text{if} \quad x < 0 \end{cases}$$

- Generalization:

$$v \in \partial \|x\|_1 \iff v_i = \begin{cases} v_i = \mathbf{sign}(x_i) & \text{if} \quad x_i \neq 0 \\ v_i \in [-1, 1] & \text{if} \quad x_i = 0 \end{cases}$$

**Proof.** We have, by applying the calculus rules

$$\|x\|_1 = \sum |x_i| = \sum |e_i^\top x|$$

hence

$$\partial \|x\|_1 = \sum \partial |e_i^\top x| = \sum e_i \partial |x_i|$$

which leads to the desired result.

□

Let $f : \mathbb{R}^n \to \mathbb{R}$. By definition of the conjugate function

$$\forall x, y \in \mathbb{R}^n, \quad x^\top y \leqslant f(x) + f^*(y)$$

Equality occurs when $y \in \partial f(x)$, *i.e.*

$$\forall x, y \in \mathbb{R}^n, \quad x^\top y = f(x) + f^*(y) \iff y \in \partial f(x)$$

**Proof.** We have

$$\begin{aligned} x^\top y \geqslant f(x) + f^*(y) &\iff x^\top y \geqslant f(x) + z^\top y - f(z) \quad \forall z \in \mathbb{R}^n \\ &\iff f(z) \geqslant f(x) + \langle y, z - x \rangle \quad \forall z \in \mathbb{R}^n \\ &\iff y \in \partial f(x) \end{aligned}$$

- The Lasso objective

$$f(x) := \frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_1 \tag{Lasso}$$

is not always strictly convex: it can have several minimizers. This is in constrast for instance with the Tikhonov regularization

$$\frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_2^2$$

which is strictly convex and always admits a unique minimizer when $\lambda > 0$.

- The Lasso objective

$$f(x) := \frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_1 \tag{Lasso}$$

is not always strictly convex: it can have several minimizers. This is in constrast for instance with the Tikhonov regularization

$$\frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_2^2$$

which is strictly convex and always admits a unique minimizer when $\lambda > 0$.

- We can derive optimality conditions for Lasso

$$0 \in \partial f(x) = A^\top(Ax - y) + \lambda\partial\|x\|_1$$

***Proposition (Lasso optimality).*** $x_*$ is a minimizer of (Lasso) if and only if there exists $\eta \in \mathbb{R}^n$ such that

$$A^\top(Ax^* - y) + \lambda\eta = 0$$

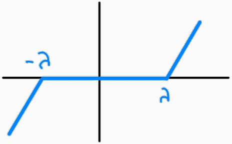
where

$$\begin{cases} \eta_i = \mathbf{sign}(x_{*i}) & \text{if} \quad x_{*i} \neq 0 \\ \eta_i \in [-1, 1] & \text{if} \quad x_{*i} = 0 \end{cases}$$

# The Othonormal Case

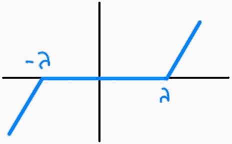
- If $A$ satisfies $A^\top A = I$, there is a closed-form solution given by the soft thresholding operator

$$S_\lambda(x) = \begin{cases} x_i + \lambda & \text{if} \quad x_i < -\lambda \\ 0 & \text{if} \quad |x_i| \leqslant \lambda \\ x_i - \lambda & \text{if} \quad x_i > \lambda \end{cases}$$



- If $A$ satisfies $A^\top A = I$, there is a closed-form solution given by the soft thresholding operator

$$S_\lambda(x) = \begin{cases} x_i + \lambda & \text{if} \quad x_i < -\lambda \\ 0 \quad \text{if} & |x_i| \leqslant \lambda \\ x_i - \lambda & \text{if} \quad x_i > \lambda \end{cases}$$



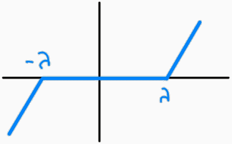

- In that case

$$\min \frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_1 = \frac{1}{2}\sum_i (x_i - (A^\top y)_i)^2 + \lambda \sum_i |x_i|,$$

so we may solve the minimization component by component (separable problem).

- If $A$ satisfies $A^\top A = I$, there is a closed-form solution given by the soft thresholding operator

$$S_\lambda(x) = \begin{cases} x_i + \lambda & \text{if} \quad x_i < -\lambda \\ 0 & \text{if} \quad |x_i| \leqslant \lambda \\ x_i - \lambda & \text{if} \quad x_i > \lambda \end{cases}$$




- In that case

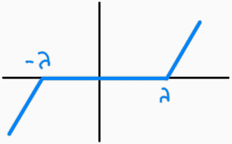
$$\min \frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_1 = \frac{1}{2}\sum_i (x_i - (A^\top y)_i)^2 + \lambda \sum_i |x_i|,$$

so we may solve the minimization component by component (separable problem). Let $z := A^\top y$ and $h : \mathbb{R} \to \mathbb{R}$, $h(x) := \frac{1}{2}(x - z)^2 + \lambda|x|$. Then the optimality conditions give

$$0 \in \partial h(x) = \begin{cases} x - z - \lambda & \text{if} \quad x < 0 \\ -z + \lambda[-1, 1] & \text{if} \quad x = 0 \\ x - z + \lambda & \text{if} \quad x > 0 \end{cases} \iff \begin{cases} x = z + \lambda & \text{if} \quad z < -\lambda \\ x = 0 & \text{if} \quad -\lambda \leqslant z \leqslant \lambda \\ x = z - \lambda & \text{if} \quad z > \lambda \end{cases}$$

- If $A$ satisfies $A^\top A = I$, there is a closed-form solution given by the soft thresholding operator

$$S_\lambda(x) = \begin{cases} x_i + \lambda & \text{if} \quad x_i < -\lambda \\ 0 & \text{if} \quad |x_i| \leqslant \lambda \\ x_i - \lambda & \text{if} \quad x_i > \lambda \end{cases}$$




- In that case

$$\min \frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_1 = \frac{1}{2}\sum_i (x_i - (A^\top y)_i)^2 + \lambda \sum_i |x_i|,$$

so we may solve the minimization component by component (separable problem). Let $z := A^\top y$ and $h : \mathbb{R} \to \mathbb{R}$, $h(x) := \frac{1}{2}(x - z)^2 + \lambda|x|$. Then the optimality conditions give

$$0 \in \partial h(x) = \begin{cases} x - z - \lambda & \text{if} \quad x < 0 \\ -z + \lambda[-1, 1] & \text{if} \quad x = 0 \\ x - z + \lambda & \text{if} \quad x > 0 \end{cases} \iff \begin{cases} x = z + \lambda & \text{if} \quad z < -\lambda \\ x = 0 & \text{if} \quad -\lambda \leqslant z \leqslant \lambda \\ x = z - \lambda & \text{if} \quad z > \lambda \end{cases}$$

- Therefore, a solution obeys $\boxed{x_* = S_\lambda(A^\top y)}$

- In general, Lasso has no closed-form solution: one must resort to iterative algorithms to approximate the solution.

- In general, LASSO has no closed-form solution: one must resort to iterative algorithms to approximate the solution.
- Gradient descent evolves in the direction of the negative gradient

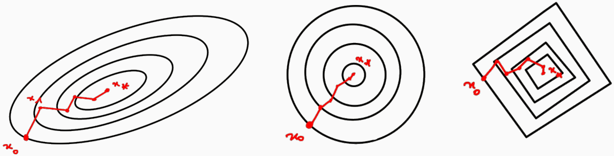
$$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

  - ✓ simple and cheap
  - ✓ can be fast for smooth (well-conditioned), strongly convex functions, with convergence at least $f(x_t) - f(x_*) = O(c^{-t})$
  - ✗ usually slow, with convergence $f(x_t) - f(x_*) = O(1/t)$
  - ✗ cannot handle non-differentiable functions

- In general, Lasso has no closed-form solution: one must resort to iterative algorithms to approximate the solution.
- Gradient descent evolves in the direction of the negative gradient

$$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

  - ✓ simple and cheap
  - ✓ can be fast for smooth (well-conditioned), strongly convex functions, with convergence at least $f(x_t) - f(x_*) = O(c^{-t})$
  - ✗ usually slow, with convergence $f(x_t) - f(x_*) = O(1/t)$
  - ✗ cannot handle non-differentiable functions



- Subgradient descent uses any vector in the subdifferential instead of the gradient

$$x_{k+1} = x_k - \gamma g_k, \quad \text{where} \quad g_k \in \partial f(x_k)$$

  - ✓ simple and cheap
  - ✗ sub-optimal solutions
  - ✗ slow, with convergence $f(x_t) - f(x_*) = O(1/\sqrt{t})$

- ***Definition (Proximal Operator).*** For a convex $f : \mathbb{R}^n \to \mathbb{R}$ (or $\overline{\mathbb{R}}$), we define its proximal operator as

$$\mathrm{prox}_{\gamma f}(x) = \mathrm{argmin}_y \, \frac{1}{2}\|x - y\|^2 + \gamma f(y)$$

- ***Definition (Proximal Operator).*** For a convex $f : \mathbb{R}^n \to \mathbb{R}$ (or $\overline{\mathbb{R}}$), we define its proximal operator as

$$\mathrm{prox}_{\gamma f}(x) = \mathrm{argmin}_y \frac{1}{2}\|x - y\|^2 + \gamma f(y)$$

- Connection with gradient descent: If $f \in C^1$, first order optimality yields

$$y = x - \gamma \nabla f(y) \ (= \mathrm{prox}_{\gamma f}(x))$$

- ***Definition (Proximal Operator).*** For a convex $f : \mathbb{R}^n \to \mathbb{R}$ (or $\overline{\mathbb{R}}$), we define its proximal operator as

$$\mathrm{prox}_{\gamma f}(x) = \mathrm{argmin}_y \frac{1}{2}\|x - y\|^2 + \gamma f(y)$$

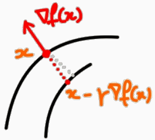
- Connection with gradient descent: If $f \in C^1$, first order optimality yields

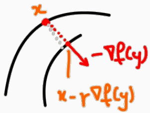
$$y = x - \gamma \nabla f(y) \; (= \mathrm{prox}_{\gamma f}(x))$$

*i.e.* y is the point from which if you look *backwards* along $-\nabla f(y)$, you reach $x$
  - Gradient step: $y - x = -\gamma \nabla f(x)$ (forward step)
  - Proximal step: $y - x = -\gamma \nabla f(y)$ (backward step)

- ***Definition (Proximal Operator).*** For a convex $f : \mathbb{R}^n \to \mathbb{R}$ (or $\overline{\mathbb{R}}$), we define its proximal operator as
$$\mathrm{prox}_{\gamma f}(x) = \mathrm{argmin}_y \frac{1}{2}\|x-y\|^2 + \gamma f(y)$$

- Connection with gradient descent: If $f \in C^1$, first order optimality yields
$$y = x - \gamma \nabla f(y) \ (= \mathrm{prox}_{\gamma f}(x))$$
*i.e.* y is the point from which if you look *backwards* along $-\nabla f(y)$, you reach $x$
  - Gradient step: $y - x = -\gamma \nabla f(x)$ (forward step)
  - Proximal step: $y - x = -\gamma \nabla f(y)$ (backward step)




If $f \in C^0$, then
$$0 \in \gamma \partial f(y) + (y - x) \iff x \in (I + \gamma \partial f)(y)$$

- ***Definition (Proximal Operator).*** For a convex $f : \mathbb{R}^n \to \mathbb{R}$ (or $\overline{\mathbb{R}}$), we define its proximal operator as

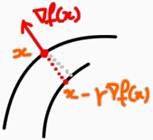
$$\mathrm{prox}_{\gamma f}(x) = \mathrm{argmin}_y \frac{1}{2}\|x - y\|^2 + \gamma f(y)$$

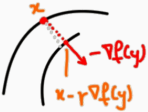
- Connection with gradient descent: If $f \in C^1$, first order optimality yields

$$y = x - \gamma \nabla f(y) \ (= \mathrm{prox}_{\gamma f}(x))$$

*i.e.* y is the point from which if you look *backwards* along $-\nabla f(y)$, you reach $x$

  - Gradient step: $y - x = -\gamma \nabla f(x)$ (forward step)
  - Proximal step: $y - x = -\gamma \nabla f(y)$ (backward step)




If $f \in C^0$, then

$$0 \in \gamma \partial f(y) + (y - x) \iff x \in (I + \gamma \partial f)(y)$$

- Proximal operator generalizes projection: if $f(x) = \mathbb{1}_C(x)$ is an indicator function of a convex set, then $\mathrm{prox}_{\gamma f}(x) = \mathrm{Proj}_C(x)$. More generally, $\mathrm{prox}_{\gamma f}(x)$ is an orthogonal projection on a level set of $f$.

- ***Proposition (Fixed point).*** Let $f : \mathbb{R}^n \to \mathbb{R}$ (or $\overline{\mathbb{R}}$) be continuous convex. For *any* $\gamma > 0$,

$$x_* \in \operatorname{argmin} f(x) \iff x_* = \operatorname{prox}_{\gamma f}(x_*)$$

**Proof.** We can assume without loss of generality that $\gamma = 1$. Suppose $f(x) \geqslant f(x_*)$ for all $x$. Then

$$\begin{aligned} f(x) + \frac{1}{2}\|x - x_*\|^2 \geqslant f(x_*) + \frac{1}{2}\|x_* - x_*\|^2 &\implies x_* = \operatorname{argmin}_x f(x) + \frac{1}{2}\|x - x_*\|^2 \\ &\implies x_* = \operatorname{prox}_f(x_*) \end{aligned}$$

On the other hand, assume that $x_* = \operatorname{prox}_f(x_*)$. Then

$$0 \in \partial f(x_*) + x_* - x_* \implies 0 \in \partial f(x_*)$$

which shows that $x_*$ minimizes $f$.

□

- ***Proposition (Fixed point).*** Let $f : \mathbb{R}^n \to \mathbb{R}$ (or $\overline{\mathbb{R}}$) be continuous convex. For *any* $\gamma > 0$,

$$x_* \in \operatorname{argmin} f(x) \iff x_* = \operatorname{prox}_{\gamma f}(x_*)$$

**Proof.** We can assume without loss of generality that $\gamma = 1$. Suppose $f(x) \geqslant f(x_*)$ for all $x$. Then

$$\begin{aligned} f(x) + \frac{1}{2}\|x - x_*\|^2 \geqslant f(x_*) + \frac{1}{2}\|x_* - x_*\|^2 &\implies x_* = \operatorname{argmin}_x f(x) + \frac{1}{2}\|x - x_*\|^2 \\ &\implies x_* = \operatorname{prox}_f(x_*) \end{aligned}$$

On the other hand, assume that $x_* = \operatorname{prox}_f(x_*)$. Then

$$0 \in \partial f(x_*) + x_* - x_* \implies 0 \in \partial f(x_*)$$

which shows that $x_*$ minimizes $f$.

□

- Proximal iterations: $x_{k+1} = \operatorname{prox}_{\gamma f}(x_k)$ (fixed-point iterations)

  *Remark.* prox is usually not a contraction (contraction = $\|h(x) - h(y)\| \leqslant \rho\|x - y\|$ with $\rho < 1$), but it is nonexpansive, and slightly more, which ensures the convergence of fixed point iterations.

- Consider the generic problem

$$\min F(x) + G(x)$$

where $F \in C^1$ is differentiable, and $G \in C^0$ is "proximable" (*i.e.* we can easily project on its level lines).

- Consider the generic problem

$$\min F(x) + G(x)$$

where $F \in C^1$ is differentiable, and $G \in C^0$ is "proximable" (*i.e.* we can easily project on its level lines). Then

$$\begin{aligned} 0 \in \nabla F(x) + \partial G(x) &\iff 0 \in (\lambda \nabla F(x) - x) + (x - \partial G(x)) \\ &\iff (I - \gamma \nabla F)(x) \in (I + \gamma \partial G)(x) \end{aligned}$$

- Suggests updates of the form

$$x_{k+1} = \mathbf{prox}_{\gamma F}(x_k - \gamma \nabla F(x_k))$$

This algorithm is called proximal gradient method, or forward-backward splitting.

# Proximal Splitting

- Consider the generic problem
$$\min F(x) + G(x)$$
where $F \in C^1$ is differentiable, and $G \in C^0$ is "proximable" (*i.e.* we can easily project on its level lines). Then
$$\begin{aligned} 0 \in \nabla F(x) + \partial G(x) &\iff 0 \in (\lambda \nabla F(x) - x) + (x - \partial G(x)) \\ &\iff (I - \gamma \nabla F)(x) \in (I + \gamma \partial G)(x) \end{aligned}$$

- Suggests updates of the form
$$x_{k+1} = \mathbf{prox}_{\gamma F}(x_k - \gamma \nabla F(x_k))$$
This algorithm is called proximal gradient method, or forward-backward splitting.

- Convergence with rate $O(1/k)$ when $\gamma \in [0, 1/L]$ fixed, where $\nabla F$ is $L$-Lipschitz ($L$ corresponds to the conditioning of $A$ in our case).

- Recall the Lasso

$$\min \frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_1 = \text{"smooth"} + \text{"simple"}$$

- Recall the LASSO

$$\min \frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_1 = \text{"smooth"} + \text{"simple"}$$

- Proximal operator for $\|\cdot\|_1$

$$\mathrm{prox}_{\lambda\|\cdot\|_1}(x) = \mathrm{argmin}_z \frac{1}{2}\|x - z\|^2 + \lambda\|z\|_1$$

We have already seen that $\mathrm{prox}_{\lambda\|\cdot\|_1} = S_\lambda(z)$

- Recall the LASSO

$$\min \frac{1}{2}\|Ax - y\|^2 + \lambda\|x\|_1 = \text{"smooth"} + \text{"simple"}$$

- Proximal operator for $\|\cdot\|_1$

$$\mathbf{prox}_{\lambda\|\cdot\|_1}(x) = \mathbf{argmin}_z \frac{1}{2}\|x - z\|^2 + \lambda\|z\|_1$$

We have already seen that $\mathbf{prox}_{\lambda\|\cdot\|_1} = S_\lambda(z)$

- Iterative Soft-thresholding Algorithm (ISTA)

$$x_{k+1} = S_\lambda\left(x_k - \frac{1}{\kappa(A)}A^\top(Ax_k - y)\right)$$

- Choice of regularization parameter $\lambda$ is, as always, sensitive

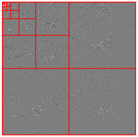
# Wavelet Sparsity

- Wavelet basis = orthonormal (Hilbert) basis of $L^2(\Omega)$, and a fortiori of $\mathbb{R}^n$

$$\psi_{a,b}^{(\theta)}(x) = \frac{1}{\sqrt{a}}\psi^{(\theta)}\left(\frac{x-b}{a}\right)$$

Comparable to Fourier basis, but extracts both spatial and frequency information. Images have sparse representation with respect to wavelets, *i.e.* $\langle f, \psi_{a,b} \rangle \simeq 0$ often.

$x \in \mathbb{R}^n$ coefficients | $f = \Psi x \in \mathbb{R}^q$ image | $y = Kf + \delta \in \mathbb{R}^m$ data



$A = K \circ \Psi \in \mathbb{R}^{m \times n}$

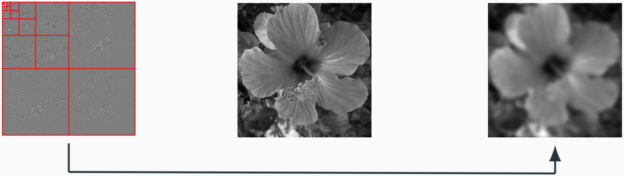
# Wavelet Sparsity

- Wavelet basis = orthonormal (Hilbert) basis of $\mathrm{L}^2(\Omega)$, and a fortiori of $\mathbb{R}^n$

$$\psi_{a,b}^{(\theta)}(x) = \frac{1}{\sqrt{a}}\psi^{(\theta)}\left(\frac{x-b}{a}\right)$$

Comparable to Fourier basis, but extracts both spatial and frequency information. Images have sparse representation with respect to wavelets, *i.e.* $\langle f, \psi_{a,b}\rangle \simeq 0$ often.

$x \in \mathbb{R}^n$ coefficients $\quad f = \boldsymbol{\Psi} x \in \mathbb{R}^q$ image $\quad y = Kf + \delta \in \mathbb{R}^m$ data



$$A = K \circ \boldsymbol{\Psi} \in \mathbb{R}^{m \times n}$$

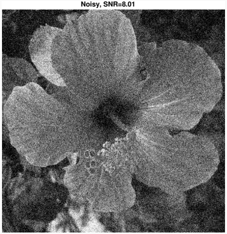
- Wavelet Sparse Regularization

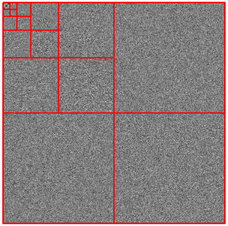
$$\boxed{\min_{x \in \mathbb{R}^n} \|y - Ax\|^2 + \|x\|_1} \qquad \text{(synthesis)}$$

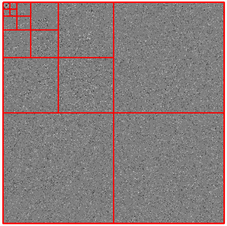
and the reconstructed image is then given by $f = \boldsymbol{\Psi} x$, or

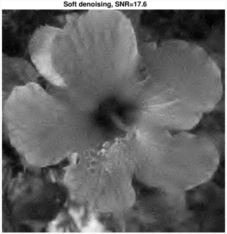
$$\boxed{\min_{f \in \mathbb{R}^q} \|y - Kf\|^2 + \|\boldsymbol{\Psi}^\top f\|_1} \qquad \text{(analysis)}$$

Corresponds to $A = I_n$: solution is given in closed-form by soft-thresholding.

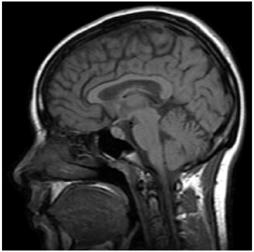






$$\min_f \|y - \mathcal{R}f\|^2 + \|\Psi^\top f\|, \quad \text{where} \quad \mathcal{R}f(s,u) = \int_{\mathbb{R}} f(su + tu^\top)\mathrm{d}t$$



Original


Pseudo-inverse


ISTA

# TV regularization

- 1D discrete total variation

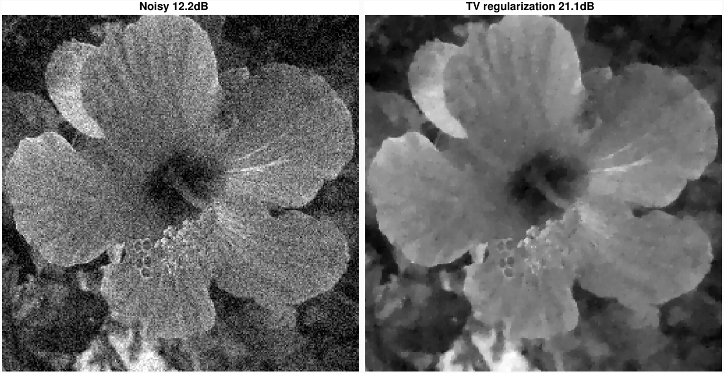
$$\min_x \|Ax - y\|^2 + \lambda \|Dx\|_1 \quad \text{where} \quad D = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & & \\ & & & 1 & -1 \end{bmatrix} \in \mathbb{R}^{(n-1)\times n}$$

Penalizes "edges" in $x$, tends to produce results piecewise constant (sparse gradient)

- nD, continuous (infinite dimensional): for smooth $f$,

$$\min \| \int K(s,t) f(t) \mathrm{d}t - y(s)\|_{L^2}^2 + \lambda \|\nabla f\|_1$$

$J(f) = \|\nabla f\|_1$ is the total variation of $f$, and it can be extended to non-smooth images with discontinuities (edges).
$J(f)$ corresponds to the total length of its level sets.
Difficult to minimize: $\nabla J(f) = \mathrm{div}(\nabla f / \|\nabla f\|)$ is not well defined everywhere.


Noisy 12.2dB
TV regularization 21.1dB

- The proximal operator for $\mathrm{TV} : x \mapsto \|Dx\|_1$ has no closed-form

$$x = \mathrm{prox}_{\gamma \mathrm{TV}}(z) \iff z \in S^\top x + \gamma\, \mathbf{sign}(Dx)$$

where $S = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ & & & \\ 1 & \dots & 1 & 1 \end{bmatrix}$ (it is the matrix such that $DS = I$)]

- Alternating Directions Method of Multipliers considers the augmented Lagrangian

$$\mathcal{L}(x, z, \nu) = \frac{1}{2}\|Ax - y\|^2 + \lambda\|z\|_1 + \nu^\top (Dx - z) + \frac{\rho}{2}\|Dx - z\|^2$$

  and solves it iteratively minimizing over $x$ (proximal step), $z$ (proximal step) and maximizing over $\nu$ (gradient ascent).

- Alternating Directions Method of Multipliers considers the augmented Lagrangian

$$\mathcal{L}(x, z, \nu) = \frac{1}{2}\|Ax - y\|^2 + \lambda\|z\|_1 + \nu^\top(Dx - z) + \frac{\rho}{2}\|Dx - z\|^2$$

  and solves it iteratively minimizing over $x$ (proximal step), $z$ (proximal step) and maximizing over $\nu$ (gradient ascent).
- Majorization-Minimization algorithm compute s at each step a (quadratic) majorant of TV($x$), and minimize it