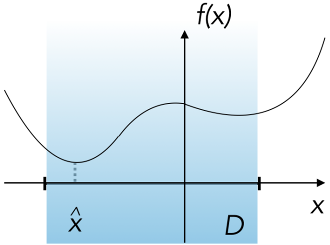
Optimization
Part I: Introduction

Nelly Pustelnik

CNRS, Laboratoire de Physique de l'ENS de Lyon, France

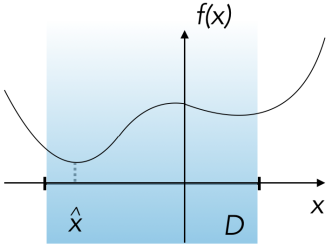
Minimization problems

- **Minimization problems** involving :
  - a cost function $f : \mathbb{R}^N \to \mathbb{R}$;
  - a subset D of $\mathbb{R}^N$.



Minimization problems

- **Minimization problems** involving :
  - a cost function $f : \mathbb{R}^N \to \mathbb{R}$;
  - a subset D of $\mathbb{R}^N$.

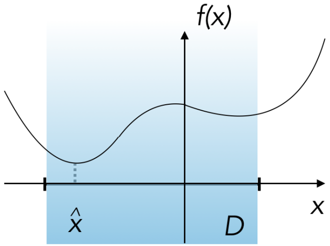


- Goal: We want to

$$
\begin{aligned}
& \text{Find } \widehat{x} \in D \quad \text{such that} \quad (\forall x \in D)\ f(\widehat{x}) \leq f(x) \\
\Leftrightarrow \quad & \text{Find } \widehat{x} \in D \quad \text{such that} \quad f(\widehat{x}) = \inf_{x \in D} f(x) \\
\Leftrightarrow \quad & \text{Find } \widehat{x} \in \operatorname*{Argmin}_{x \in D} \ f(x).
\end{aligned}
$$

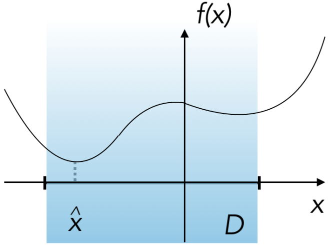
Minimization problems

- **Maximization problems** involving :
  - a reward function $f : \mathbb{R}^N \to \mathbb{R}$;
  - a subset D of $\mathbb{R}^N$.



Minimization problems

- **Maximization problems** involving :
  - a reward function $f : \mathbb{R}^N \to \mathbb{R}$;
  - a subset D of $\mathbb{R}^N$.



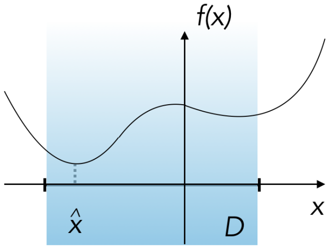
- Goal: We want to

$$
\begin{aligned}
& \text{Find } \widehat{x} \in D \quad \text{such that} \quad (\forall x \in D)\ f(\widehat{x}) \geq f(x) \\
\Leftrightarrow \quad & \text{Find } \widehat{x} \in D \quad \text{such that} \quad (\forall x \in D)\ -f(\widehat{x}) \leq -f(x) \\
\Leftrightarrow \quad & \text{Find } \widehat{x} \in \underset{x \in D}{\operatorname{Argmin}} \ (-f(x)) .
\end{aligned}
$$

Minimization problems

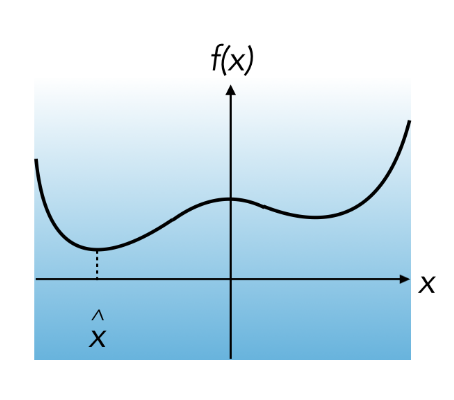
- **Maximization problems** involving :
  - a reward function $f : \mathbb{R}^N \to \mathbb{R}$;
  - a subset D of $\mathbb{R}^N$.

Without loss of generality, we can focus on minimization problems



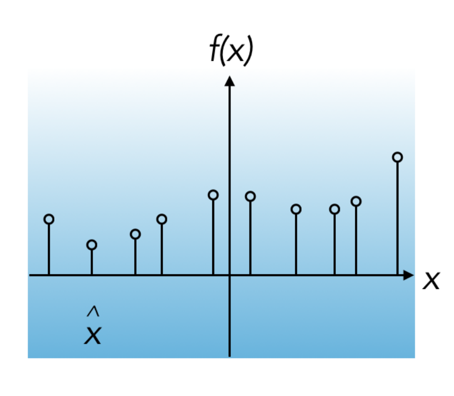
Various types of minimization problems

- $D = \mathbb{R}^N$: **unconstrained problem**



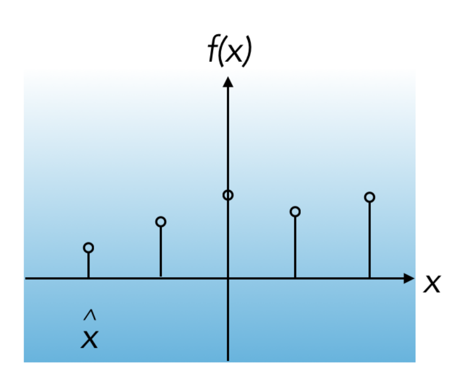
Various types of minimization problems

- D **countable**: discrete optimization problem



Various types of minimization problems

- D **countable**: discrete optimization problem



Various types of minimization problems

- D **uncountable**: continuous optimization problem

  - Example: Optimization problem with P equality constraints

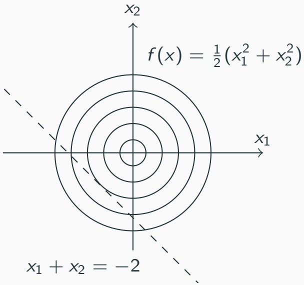
$$D = \big\{x \in \mathbb{R}^N \mid (\forall i \in \{1, \dots, P\}) \quad g_i(x) = 0\big\}$$

where $g_i : \mathbb{R}^N \to \mathbb{R}$.

  - Particular case: linear (or affine) constraints

$$\begin{aligned} g_i(x) &= \langle a_i \mid x \rangle + b_i \\ &= \sum_{n=1}^{N} a_{i,n} x_n + b_i \end{aligned}$$

where $a_i \in \mathbb{R}^N$ and $\mathrm{b}_i \in \mathbb{R}$.



Various types of minimization problems

- D **uncountable**: continuous optimization problem

  - Example: Optimization problem with P inequality constraints

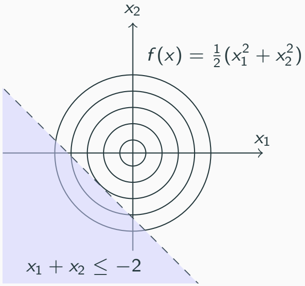
$$D = \{x \in \mathbb{R}^N \mid (\forall i \in \{1, \ldots, P\}) \quad g_i(x) \leq 0\}$$

where $g_i : \mathbb{R}^N \to \mathbb{R}$.

  - Particular case: linear (or affine) constraints

$$g_i(x) = \langle a_i \mid x \rangle + b_i$$
$$= \sum_{n=1}^{N} a_{i,n} x_n + b_i$$

where $a_i \in \mathbb{R}^N$ and $b_i \in \mathbb{R}$.



Constrained and unconstrained minimization problems

- Reformulation using **indicator function**

$$\text{Find } \widehat{x} \in \underset{x \in D}{\operatorname{Argmin}} \; f(x) \quad \Leftrightarrow \quad \text{Find } \widehat{x} \in \underset{x \in \mathbb{R}^N}{\operatorname{Argmin}} \; f(x) + \iota_D(x)$$

where

$$(\forall x \in \mathbb{R}^N) \quad \iota_D(x) = \begin{cases} 0 & \text{if } x \in D \\ +\infty & \text{otherwise.} \end{cases}$$

Constrained and unconstrained minimization problems

- Reformulation using **indicator function**

$$\text{Find } \widehat{x} \in \underset{x \in D}{\operatorname{Argmin}} \; f(x) \quad \Leftrightarrow \quad \text{Find } \widehat{x} \in \underset{x \in \mathbb{R}^N}{\operatorname{Argmin}} \; f(x) + \iota_D(x)$$

where

$$(\forall x \in \mathbb{R}^N) \quad \iota_D(x) = \begin{cases} 0 & \text{if } x \in D \\ +\infty & \text{otherwise.} \end{cases}$$

- or equivalently

$$\text{Find } \widehat{x} \in \underset{x \in \mathbb{R}^N}{\operatorname{Argmin}} \; \widetilde{f}(x)$$

where

$$(\forall x \in \mathbb{R}^N) \quad \widetilde{f}(x) = \begin{cases} f(x) & \text{if } x \in D \\ +\infty & \text{otherwise.} \end{cases}$$

Constrained and unconstrained minimization problems

- Reformulation using **indicator function**

$$\text{Find } \widehat{x} \in \underset{x \in D}{\mathrm{Argmin}} \; f(x) \quad \Leftrightarrow \quad \text{Find } \widehat{x} \in \underset{x \in \mathbb{R}^N}{\mathrm{Argmin}} \; f(x) + \iota_D(x)$$

where

$$(\forall x \in \mathbb{R}^N) \quad \iota_D(x) = \begin{cases} 0 & \text{if } x \in D \\ +\infty & \text{otherwise.} \end{cases}$$

Allowing non finite valued functions leads to a unifying view of constrained and unconstrained minimization problems.

Main questions to be addressed

1. **Existence/uniqueness** of a solution $\widehat{x}$?

Main questions to be addressed

1. **Existence/uniqueness** of a solution $\widehat{x}$?
2. Characterization of solutions: **necessary/sufficient conditions** for $\widehat{x}$ to be a solution.

Main questions to be addressed

1. **Existence/uniqueness** of a solution $\widehat{x}$?
2. Characterization of solutions: **necessary/sufficient conditions** for $\widehat{x}$ to be a solution.
3. Designing an **algorithm** to approximate a solution in the frequent case when no closed form solution is available, i.e. building a sequence $(x_n)_{n\in\mathbb{N}}$ of $\mathbb{R}^N$ such that

$$\lim_{n\to+\infty} x_n = \widehat{x}.$$

Main questions to be addressed

1. **Existence/uniqueness** of a solution $\widehat{x}$?
2. Characterization of solutions: **necessary/sufficient conditions** for $\widehat{x}$ to be a solution.
3. Designing an **algorithm** to approximate a solution in the frequent case when no closed form solution is available, i.e. building a sequence $(x_n)_{n\in\mathbb{N}}$ of $\mathbb{R}^N$ such that

$$\lim_{n\to+\infty} x_n = \widehat{x}.$$

4. Evaluation of the performance of the optimization algorithm:
   - **Convergence rate**
     Example: If there exists $\rho\in]0,1[$ and $n^*\in\mathbb{N}$ such that $(\forall n\geq n^*)$
     $\|x_{n+1}-\widehat{x}\|\leq\rho\|x_n-\widehat{x}\|$, then $(Q\text{-})$*linear* convergence rate.
     If $\lim_{n\to+\infty}\frac{\|x_{n+1}-\widehat{x}\|}{\|x_n-\widehat{x}\|}=0$, then *superlinear* convergence rate.
     $\|x_{n+1}-\widehat{x}\|\leq\rho\|x_n-\widehat{x}\|^2$, then quadratic convergence rate.
   - **Robustness** to numerical errors
   - Amenability to **parallel/distributed implementations**.

Example: Supervised learning

- **Database**: $\mathcal{S} = \big\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\big|\; \ell \in \{1, \ldots, L\}\big\}$
  e.g. $u_\ell \in \underbrace{\mathbb{R}^N}_{\mathcal{H}}$ image and $z_\ell \in \underbrace{\{-1, 1\}}_{\mathcal{G}}$ classe (chameleon/stick insect)

Example: Supervised learning

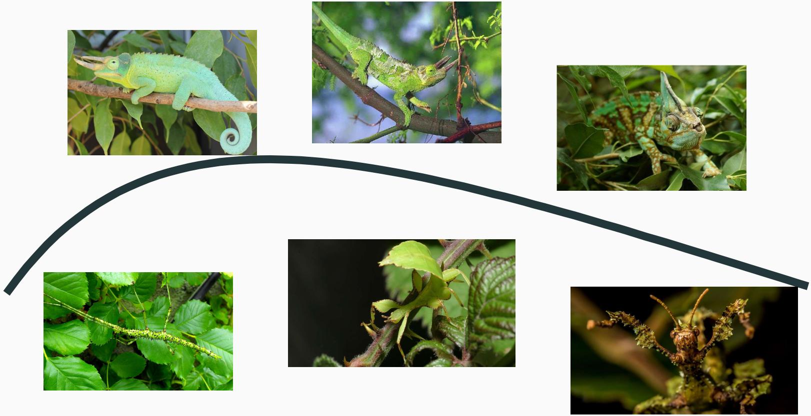
- **Database**: $\mathcal{S} = \left\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\middle|\; \ell \in \{1, \ldots, L\}\right\}$
  e.g. $u_\ell \in \underbrace{\mathbb{R}^N}_{\mathcal{H}}$ image and $z_\ell \in \underbrace{\{-1, 1\}}_{\mathcal{G}}$ classe (chameleon/stick insect)

- **Goal**: Learn a prediction function $d \colon \mathcal{H} \to \mathcal{G}$

Example: Supervised learning

- **Database**: $\mathcal{S} = \big\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\big|\; \ell \in \{1, \dots, L\}\big\}$
  e.g. $u_\ell \in \underbrace{\mathbb{R}^N}_{\mathcal{H}}$ image and $z_\ell \in \underbrace{\{-1, 1\}}_{\mathcal{G}}$ classe (chameleon/stick insect)

- **Goal**: Learn a prediction function $d \colon \mathcal{H} \to \mathcal{G}$



Example: Supervised learning

- **Database**: $\mathcal{S} = \{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \mid \ell \in \{1, \ldots, L\}\}$

- **Goal**: Learn a prediction function $d \colon \mathcal{H} \to \mathcal{G}$

- **Typical choices for $\mathcal{H}$**:
  - $\mathcal{H} = \mathbb{R}^N$ for univariate signal with N samples;
  - $\mathcal{H} = \mathbb{R}^{N \times M}$ for multivariate signal with N samples and M components;
  - $\mathcal{H} = \mathbb{R}^N$ for image of size $N = N_1 \times N_2$;
  - $\mathcal{H} = \mathbb{R}^{N \times M}$ for graphs with N nodes and a multiv. information on each node.

Example: Supervised learning

- **Database**: $\mathcal{S} = \big\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\big|\; \ell \in \{1, \ldots, L\}\big\}$

- **Goal**: Learn a prediction function $d \colon \mathcal{H} \to \mathcal{G}$

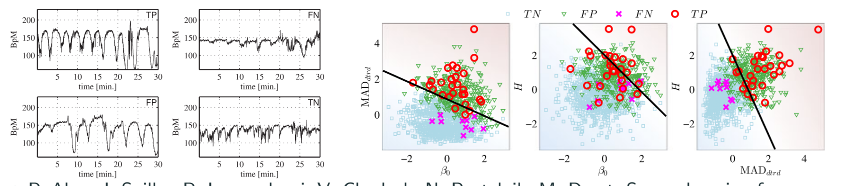
- **Typical choices for $\mathcal{G}$**:
  - $\mathcal{G} = \{-1, +1\}$ for binary classification;
  - $\mathcal{G} = \{1, \ldots, K\}$ for multiclass classification;
  - $\mathcal{G} = \mathbb{R}$ for regression;
  - $\mathcal{G} = \mathbb{R}^K$ for multivariate regression;

Example: Supervised learning

- **Database**: $\mathcal{S} = \big\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\big|\; \ell \in \{1, \ldots, L\}\big\}$

- **Goal**: Learn a prediction function $d\colon \mathcal{H} \to \mathcal{G}$

- **Typical choices for $\mathcal{G}$**:
  - $\mathcal{G} = \{-1, +1\}$ for binary classification;
  - $\mathcal{G} = \{1, \ldots, K\}$ for multiclass classification;
  - $\mathcal{G} = \mathbb{R}$ for regression;
  - $\mathcal{G} = \mathbb{R}^K$ for multivariate regression;



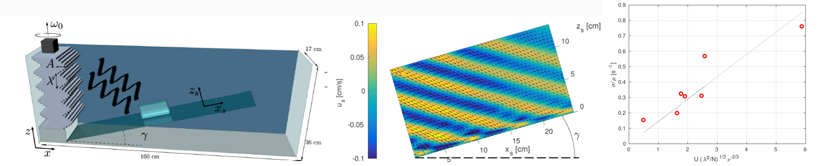
→ P. Abry, J. Spilka, R. Leonarduzzi, V. Chudcek, N. Pustelnik, M. Doret, Sparse learning for Intrapartum fetal heart rate analysis, Biomed. Phys. & Eng. Express, 4(3):034002, 2018. [PDF]

Example: Supervised learning

- **Database**: $\mathcal{S} = \{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \mid \ell \in \{1, \ldots, L\}\}$

- **Goal**: Learn a prediction function $d \colon \mathcal{H} \to \mathcal{G}$

- **Typical choices for $\mathcal{G}$**:
  - $\mathcal{G} = \{-1, +1\}$ for binary classification;
  - $\mathcal{G} = \{1, \ldots, K\}$ for multiclass classification;
  - $\mathcal{G} = \mathbb{R}$ for regression;
  - $\mathcal{G} = \mathbb{R}^K$ for multivariate regression;



→ J. Schmitt, E. Horne, N. Pustelnik, S. Joubaud, P. Odier An improved variational mode decomposition method for internal waves separation, EUSIPCO, Nice, France, 2015. [PDF]

Example: Supervised learning

- **Database**: $\mathcal{S} = \left\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\middle|\; \ell \in \{1, \ldots, L\}\right\}$
- **Goal**: Learn a prediction function $d \colon \mathcal{H} \to \mathcal{G}$
- Learning procedure relies on a **minimization problem**:

$$\underset{d}{\text{minimize}} \quad \underbrace{\frac{1}{L}\sum_{\ell=1}^{L} f_1(z_\ell, d(u_\ell))}_{\text{Data-term}} \;+\; \lambda \underbrace{f_2(d)}_{\text{Prior}}$$

Example: Supervised learning

- **Database**: $\mathcal{S} = \big\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\big|\; \ell \in \{1, \ldots, L\}\big\}$
- **Goal**: Learn a prediction function $d \colon \mathcal{H} \to \mathcal{G}$
- Learning procedure relies on a **minimization problem**:

$$\boxed{\underset{d}{\text{minimize}} \quad \underbrace{\frac{1}{L}\sum_{\ell=1}^{L} f_1(z_\ell, d(u_\ell))}_{\text{Data-term}} \; + \; \lambda \underbrace{f_2(d)}_{\text{Prior}}}$$

- **Linear predictor**: $d(u) = x^\top u$

  ⊙ Ridge regression: $\underset{x \in \mathcal{H}}{\text{minimize}} \;\; \frac{1}{L}\sum_{\ell=1}^{L}(z_\ell - x^\top u_\ell)^2 + \lambda \|x\|_2^2$

  ⊙ Logistic classification: $\underset{x \in \mathcal{H}}{\text{minimize}} \;\; \frac{1}{L}\sum_{\ell=1}^{L} \log\Big(1 + e^{-z_\ell x^\top u_\ell}\Big) + \lambda \|x\|_2^2$

Example: Supervised learning

- **Database**: $\mathcal{S} = \big\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\big|\; \ell \in \{1, \ldots, L\}\big\}$
- **Goal**: Learn a prediction function $d \colon \mathcal{H} \to \mathcal{G}$
- Learning procedure relies on a **minimization problem**:

$$\underset{d}{\text{minimize}} \quad \underbrace{\frac{1}{L}\sum_{\ell=1}^{L} f_1(z_\ell, d(u_\ell))}_{\text{Data-term}} \; + \; \lambda \underbrace{f_2(d)}_{\text{Prior}}$$

- **Linear predictor**: $d(u) = x^\top u$

  ⊙ Ridge regression: $\underset{x \in \mathcal{H}}{\text{minimize}} \; \frac{1}{L}\sum_{\ell=1}^{L}(z_\ell - x^\top u_\ell)^2 + \lambda \|x\|_2^2$

  ⊙ Logistic classification: $\underset{x \in \mathcal{H}}{\text{minimize}} \; \frac{1}{L}\sum_{\ell=1}^{L} \log\Big(1 + e^{-z_\ell x^\top u_\ell}\Big) + \lambda \|x\|_2^2$

⇒ Convex smooth problems

Example: Supervised learning

- **Database**: $\mathcal{S} = \left\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\middle|\; \ell \in \{1, \ldots, L\}\right\}$
- **Goal**: Learn a prediction function $d \colon \mathcal{H} \to \mathcal{G}$
- Learning procedure relies on a **minimization problem**:

$$\underset{d}{\text{minimize}} \quad \underbrace{\frac{1}{L}\sum_{\ell=1}^{L} f_1(z_\ell, d(u_\ell))}_{\text{Data-term}} \; + \; \lambda \underbrace{f_2(d)}_{\text{Prior}}$$

- **Linear predictor**: $d(u) = x^\top u$

  $\Rightarrow$ can be extended to $d(u) = x^\top \phi(u)$ (e.g. ϕ scattering transform)

  ⊙ Ridge regression: $\underset{x \in \mathcal{H}}{\text{minimize}} \; \frac{1}{L}\sum_{\ell=1}^{L}(z_\ell - x^\top u_\ell)^2 + \lambda \|x\|_2^2$

  ⊙ Logistic classification: $\underset{x \in \mathcal{H}}{\text{minimize}} \; \frac{1}{L}\sum_{\ell=1}^{L} \log\left(1 + e^{-z_\ell x^\top u_\ell}\right) + \lambda \|x\|_2^2$

Example: Supervised learning

- **Database**: $\mathcal{S} = \big\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\big|\; \ell \in \{1, \ldots, L\}\big\}$
- **Goal**: Learn a prediction function $d \colon \mathcal{H} \to \mathcal{G}$
- Learning procedure relies on a **minimization problem**:

$$\underset{d}{\text{minimize}} \quad \underbrace{\frac{1}{L}\sum_{\ell=1}^{L} f_1(z_\ell, d(u_\ell))}_{\text{Data-term}} \;+\; \lambda \underbrace{f_2(d)}_{\text{Prior}}$$

- **Linear predictor**: $d(u) = x^\top u$
  - ⊙ Sparse regression: $\underset{x\in\mathcal{H}}{\text{minimize}} \;\; \frac{1}{L}\sum_{\ell=1}^{L}(z_\ell - x^\top u_\ell)^2 + \lambda\|x\|_1$
  - ⊙ Sparse logistic classification: $\underset{x\in\mathcal{H}}{\text{minimize}} \;\; \frac{1}{L}\sum_{\ell=1}^{L}\log\Big(1 + e^{-z_\ell x^\top u_\ell}\Big) + \lambda\|x\|_1$
  - ⊙ SVM classification: $\underset{x\in\mathcal{H}}{\text{minimize}} \;\; \frac{1}{L}\sum_{\ell=1}^{L}\max\Big(0, 1 - z_\ell x^\top u_\ell\Big) + \lambda\|x\|_2^2$

Example: Supervised learning

- **Database**: $\mathcal{S} = \big\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\big|\; \ell \in \{1, \ldots, L\}\big\}$
- **Goal**: Learn a prediction function $d\colon \mathcal{H} \to \mathcal{G}$
- Learning procedure relies on a **minimization problem**:

$$\underset{d}{\text{minimize}} \quad \underbrace{\frac{1}{L}\sum_{\ell=1}^{L} f_1(z_\ell, d(u_\ell))}_{\text{Data-term}} \; + \; \lambda \underbrace{f_2(d)}_{\text{Prior}}$$

- **Linear predictor**: $d(u) = x^\top u$
  - ⊙ Sparse regression: $\underset{x\in\mathcal{H}}{\text{minimize}} \; \frac{1}{L}\sum_{\ell=1}^{L}(z_\ell - x^\top u_\ell)^2 + \lambda\|x\|_1$
  - ⊙ Sparse logistic classification: $\underset{x\in\mathcal{H}}{\text{minimize}} \; \frac{1}{L}\sum_{\ell=1}^{L}\log\Big(1 + e^{-z_\ell x^\top u_\ell}\Big) + \lambda\|x\|_1$
  - ⊙ SVM classification: $\underset{x\in\mathcal{H}}{\text{minimize}} \; \frac{1}{L}\sum_{\ell=1}^{L}\max\Big(0, 1 - z_\ell x^\top u_\ell\Big) + \lambda\|x\|_2^2$

$\Rightarrow$ **Convex non-smooth problems**

Example: Supervised learning

- **Database**: $\mathcal{S} = \big\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\big|\; \ell \in \{1, \ldots, L\}\big\}$
- **Goal**: Learn a prediction function $d \colon \mathcal{H} \to \mathcal{G}$
- Learning procedure relies on a **minimization problem**:

$$\underset{d}{\text{minimize}} \quad \underbrace{\frac{1}{L}\sum_{\ell=1}^{L} f_1(z_\ell, d(u_\ell))}_{\text{Data-term}} \;+\; \lambda \underbrace{f_2(d)}_{\text{Prior}}$$

- **Kernel-based predictor** (parametrized):

$$\underset{\alpha \in \mathbb{R}^L}{\text{minimize}} \;\; \frac{1}{L}\sum_{\ell=1}^{L} f_1\big(z_\ell, (\mathrm{K}\alpha)_\ell\big) + \lambda \alpha \mathrm{K}^2 \alpha$$

  - ⊙ Kernel: $\mathrm{K}(u_\ell, u_{\ell'}) = \langle \varphi u_\ell \mid \varphi u_{\ell'} \rangle$
  - ⊙ Solution d^*: $d^* = \sum_{\ell=1}^{L} \alpha_\ell \varphi(x_\ell)$
  - ⊙ More details here: [J. Mairal course]

Example: Supervised learning

- **Database**: $\mathcal{S} = \left\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\middle|\; \ell \in \{1, \ldots, L\}\right\}$
- **Goal**: Learn a prediction function $d\colon \mathcal{H} \to \mathcal{G}$
- Learning procedure relies on a **minimization problem**:

$$\boxed{\operatorname*{minimize}_{d} \quad \underbrace{\frac{1}{L}\sum_{\ell=1}^{L} f_1(z_\ell, d(u_\ell))}_{\text{Data-term}} \;+\; \lambda \underbrace{f_2(d)}_{\text{Prior}}}$$

- **Neural network predictor** (parametrized):

$$d(u) = \eta^{[K]}(W^{[K]}\eta^{[K-1]}(W^{[K-1]} \ldots \eta^{[2]}(W^{[2]}\eta^{[1]}(W^{[1]}u)) \ldots))$$

  - ⊙ Linear operators: $W^{[1]}, W^{[2]}, \ldots, W^{[K]}$
  - ⊙ Activation functions: $\eta^{[1]}, \eta^{[2]}, \ldots, \eta^{[K]}$

Example: Supervised learning

- **Database**: $\mathcal{S} = \big\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\big|\; \ell \in \{1, \ldots, L\}\big\}$
- **Goal**: Learn a prediction function $d \colon \mathcal{H} \to \mathcal{G}$
- Learning procedure relies on a **minimization problem**:

$$\underset{d}{\text{minimize}} \quad \underbrace{\frac{1}{L} \sum_{\ell=1}^{L} f_1(z_\ell, d(u_\ell))}_{\text{Data-term}} \;+\; \lambda \underbrace{f_2(d)}_{\text{Prior}}$$

- **Neural network predictor** (parametrized):

$$d(u) = \eta^{[K]}(W^{[K]}\eta^{[K-1]}(W^{[K-1]} \ldots \eta^{[2]}(W^{[2]}\eta^{[1]}(W^{[1]}u)) \ldots))$$

⊙ Linear operators: $W^{[1]}, W^{[2]}, \ldots, W^{[K]}$

⊙ Activation functions: $\eta^{[1]}, \eta^{[2]}, \ldots, \eta^{[K]}$

$\Rightarrow$ **Non-convex problems**

Example: Inverse problems

- **Data**: We observe data $z \in \mathbb{R}^K$ being a degraded version of an original image $\overline{x} \in \mathbb{R}^N$ such that: $z = A\overline{x} + \varepsilon$
  - $A : \mathbb{R}^{K \times N}$: denotes a linear degradation (e.g. a blur)
  - ε : denotes a noise (e.g. Gaussian, Poisson noise)

- **Goal**: Restore the degraded image i.e., find $\widehat{x}$ close to $\overline{x}$:

$$\underset{x \in \mathbb{R}^N}{\text{minimize}} \quad \underbrace{f_1(Ax; z)}_{\text{Data-term}} + \lambda \underbrace{f_2(x)}_{\text{Penalization}}$$

(a) Degraded Uniform blur 9×9 Gaussian noise (b) Inverse filtering Quadratic regularisation (c) $\Lambda = \mathrm{Id}$ (d) Λ Laplacian (e) Total variation

→ N. Pustelnik, A. Benazza-Benhayia, Y. Zheng, J.-C. Pesquet, Wavelet-based Image Deconvolution and Reconstruction, Wiley Encyclopedia of EEE, DOI: 10.1002/047134608X.W8294, Feb. 2016. [PDF]

Example: Inverse problems

- **Data**: We observe data $z \in \mathbb{R}^K$ being a degraded version of an original image $\overline{x} \in \mathbb{R}^N$ such that: z$= A\overline{x} + \varepsilon$
    - $A : \mathbb{R}^{K \times N}$: denotes a linear degradation (e.g. a blur)
    - ε : denotes a noise (e.g. Gaussian, Poisson noise)

- **Goal**: Restore the degraded image i.e., find $\widehat{x}$ close to $\bar{x}$:

$$\underset{x \in \mathbb{R}^N}{\text{minimize}} \; \underbrace{f_1(Ax; z)}_{\text{Data-term}} + \lambda \; \underbrace{f_2(x)}_{\text{Penalization}}$$

- **Examples**:

⊙ Poisson noise: Kullback-Leibler divergence

$$(\forall y \in \mathbb{R}^K) \quad f_1(y; z) = \textstyle\sum_{k=1}^K \phi(y_k)$$

$$\text{where } \phi(y_k) = \begin{cases} -z_k \ln(y_k) + \alpha y_k & \text{if } y_k > 0 \text{ and } z_k > 0 \\ \alpha y_k & \text{if } y_k \geq 0 \text{ and } z_k = 0 \\ +\infty & \text{otherwise} \end{cases}$$

⊙ Correlated Gaussian noise: square Mahanalobis distance $f_1(x; z) = \|x - z\|_W^2$

Piecewise constant/linear denoising

- **Data**: We observe data $z \in \mathbb{R}^K$ being a degraded version of an original signal $\overline{x} \in \mathbb{R}^N$ such that: $z = \overline{x} + \varepsilon$
  - $\overline{x}$: assumed piecewise constant/linear.
  - ε : denotes a noise (e.g. Gaussian, Poisson noise).

- **Goal**: Restore the degraded image i.e., find $\widehat{x}$ close to $\bar{x}$:

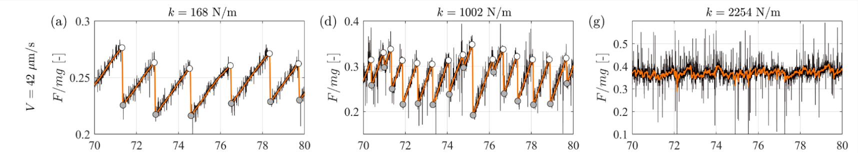
$$\underset{x \in \mathbb{R}^N}{\text{minimize}} \quad \underbrace{\frac{1}{2}\|x - z\|_2^2}_{\text{Data-term}} + \lambda \underbrace{\|Dx\|_1}_{\text{Penalization}}$$

Piecewise constant/linear denoising

- **Data**: We observe data $z \in \mathbb{R}^K$ being a degraded version of an original signal $\overline{x} \in \mathbb{R}^N$ such that: $z = \overline{x} + \varepsilon$
  - $\overline{x}$: assumed piecewise constant/linear.
  - ε : denotes a noise (e.g. Gaussian, Poisson noise).

- **Goal**: Restore the degraded image i.e., find $\widehat{x}$ close to $\bar{x}$:

$$\underset{x \in \mathbb{R}^N}{\text{minimize}} \quad \underbrace{\frac{1}{2}\|x - z\|_2^2}_{\text{Data-term}} + \lambda \underbrace{\|Dx\|_1}_{\text{Penalization}}$$

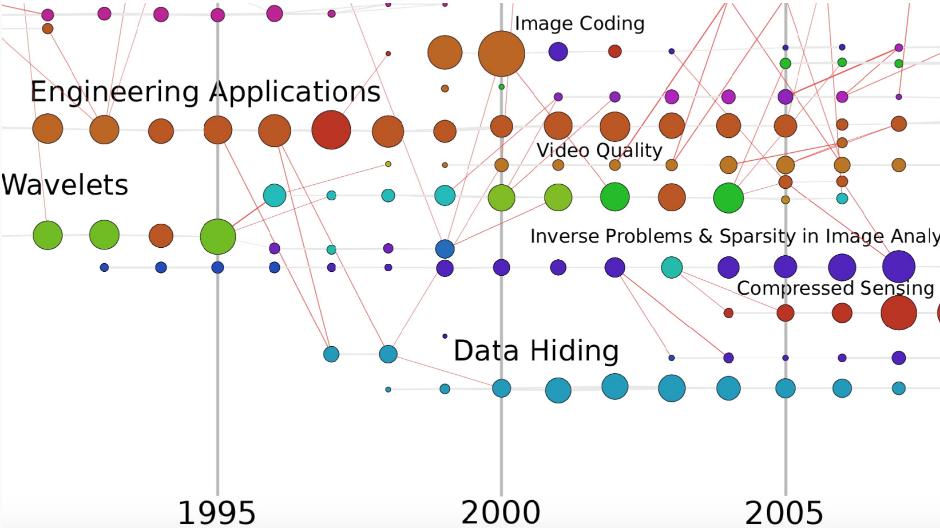


→ J. Colas, N. Pustelnik, C. Oliver, J.-C. Gminard, V. Vidal, Nonlinear denoising for solid friction dynamics characterization, Physical Review E, 100, 032803, Sept. 2019. [PDF]

Wavelet history → sparsity

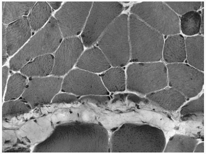
→ extracted from M. Morini, P. Flandrin, E. Fleury, T. Venturini, P. Jensen, "Revealing evolutions in dynamical networks," 2018. [PDF]

Wavelet history → sparsity

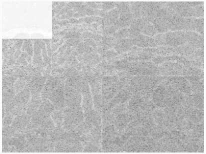


→ extracted from M. Morini, P. Flandrin, E. Fleury, T. Venturini, P. Jensen, "Revealing evolutions in dynamical networks," 2018. [PDF]

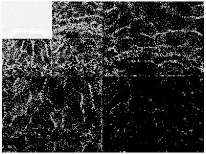
Wavelet shrinkage (Donoho-Jonhstone, 1992)



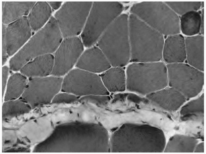
z



$\alpha = Fz$

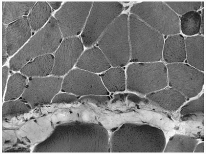


$\mathrm{soft}_\lambda(Fz)$

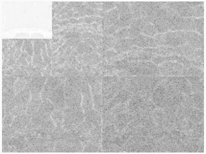


$\widehat{x} = F^* \mathrm{soft}_\lambda(Fz)$

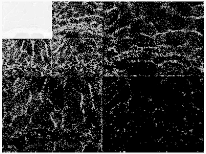
Wavelet shrinkage (Donoho-Jonhstone, 1992)



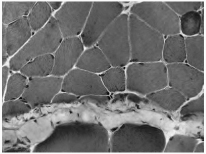
z



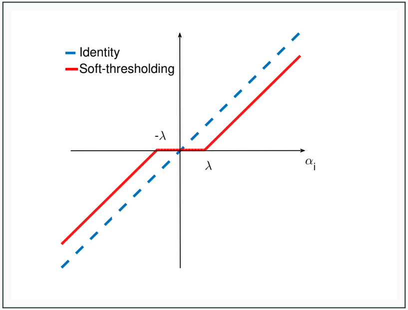
$\alpha = Fz$



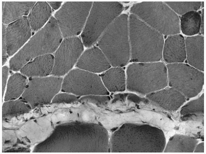
$\mathrm{soft}_\lambda(Fz)$



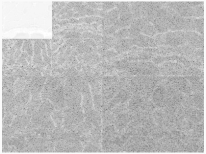
$\widehat{x} = F^*\mathrm{soft}_\lambda(Fz)$



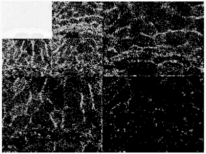
Wavelet shrinkage (Donoho-Jonhstone, 1992)



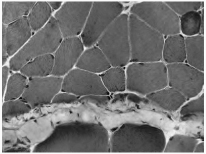
z



$\alpha = Fz$



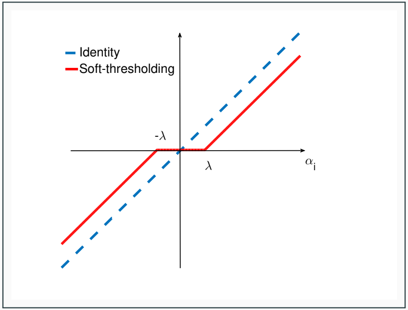
$\mathrm{soft}_\lambda(Fz)$



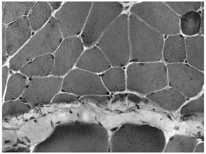
$\widehat{x} = F^* \mathrm{soft}_\lambda(Fz)$

$$\mathrm{soft}_\lambda(\alpha) = \left(\max\{|\alpha_i| - \lambda, 0\} \mathrm{sign}(\alpha_i) \right)_{1 \le i \le N}$$

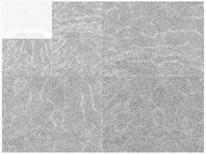
$$= \arg\min_{\nu} \frac{1}{2} \|\nu - \alpha\|_2^2 + \lambda \underbrace{\sum_i |\nu_i|}_{\|\nu\|_1}$$



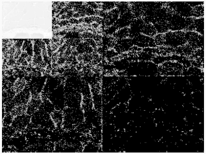
Wavelet shrinkage (Donoho-Jonhstone, 1992)



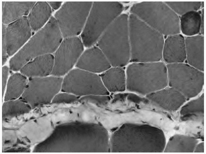
z



$\alpha = Fz$

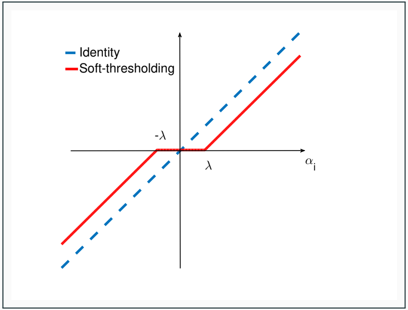


$\mathrm{soft}_\lambda(Fz)$



$\widehat{x} = F^*\mathrm{soft}_\lambda(Fz)$

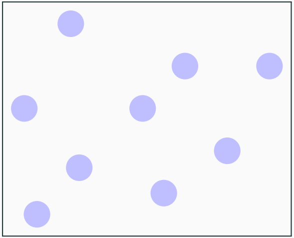
$$\begin{aligned}\mathrm{soft}_\lambda(\alpha) &= \big(\max\{|\alpha_i| - \lambda, 0\}\mathrm{sign}(\alpha_i)\big)_{1\leq i\leq N} \\ &= \arg\min_\nu \frac{1}{2}\|\nu-\alpha\|_2^2 + \lambda \underbrace{\sum_i |\nu_i|}_{\|\nu\|_1}\end{aligned}$$



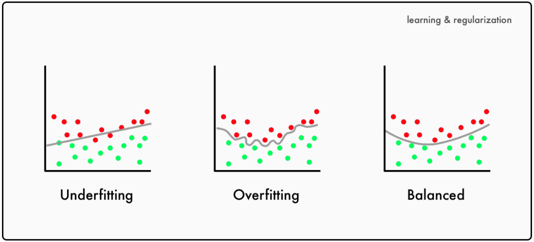
$$\widehat{x} = \arg\min_x \underbrace{\frac{1}{2}\|x-z\|_2^2}_{\text{Data-term}} + \lambda \underbrace{\|Fx\|_1}_{\text{Penalization}}$$

Sparsity in learning

(+) Perform selection

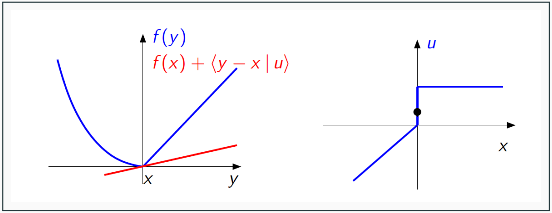


(+) Avoid overfitting



(extracted from towardsdatascience.com)

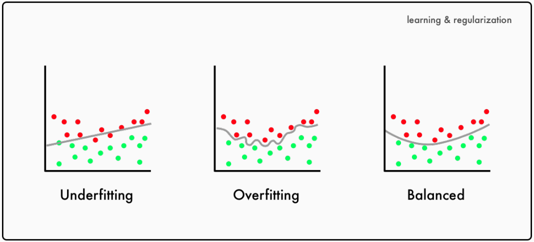
(−) Non-smooth optimization $\nabla f \Rightarrow \partial f$



Sparsity in learning

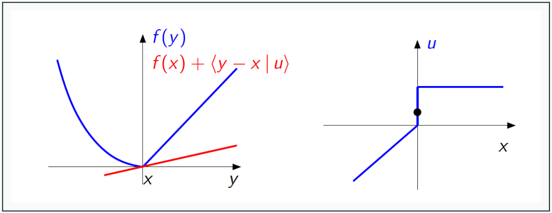
(+) Perform selection

(+) Avoid overfitting



(extracted from towardsdatascience.com)

(−) Non-smooth optimization $\nabla f \Rightarrow \partial f$



Proximal algorithms

General objective function involving linear operators L_s from $\mathbb{R}^N$ to $\mathbb{R}^{|\Upsilon_s|}$ and functions f_s proper, convex, l.s.c functions from $\mathbb{R}^{|\Upsilon_s|}$ to $]-\infty, +\infty]$:

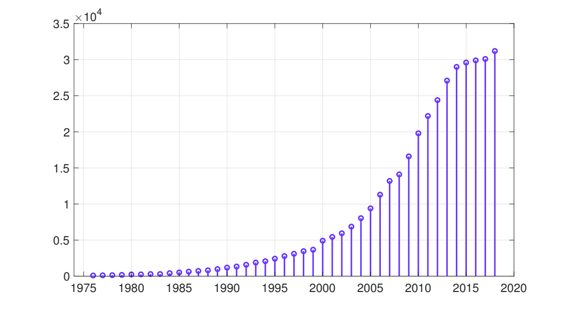
$$\widehat{x} \in \underset{x \in \mathbb{R}^N}{\operatorname{Argmin}} \sum_{s=1}^{S} f_s(L_s x)$$

- **Constraints:**
  - ⊙ Handle large datasets;
  - ⊙ Possibly non-smooth functions;
  - ⊙ Flexibility in the design of objective functions;
  - ⊙ Parallel implementation;
- **Framework:** Proximal algorithms [Bauschke-Combettes, 2017]
  - ⊙ Forward-Backward
  - ⊙ Douglas-Rachford
  - ⊙ ADMM, Primal-dual ...

Proximal algorithms

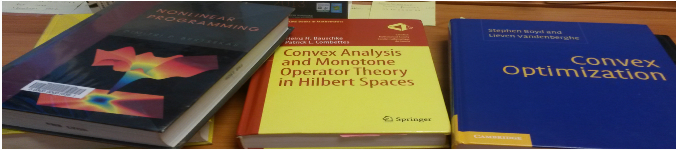
General objective function involving linear operators L_s from $\mathbb{R}^N$ to $\mathbb{R}^{|\Upsilon_s|}$ and functions f_s proper, convex, l.s.c functions from $\mathbb{R}^{|\Upsilon_s|}$ to $]-\infty,+\infty]$:

$$\widehat{x} \in \underset{x\in\mathbb{R}^N}{\mathrm{Argmin}} \sum_{s=1}^{S} f_s(L_s x)$$



Number of articles per year on Google scholar containing "proximal algorithm" since 1976.

Reference books



- **D. Bertsekas**, Nonlinear programming, Athena Scientic, Belmont, Massachussets, 1995.
- **Y. Nesterov**, Introductory Lectures on Convex Optimization: A Basic Course, Springer, 2004.
- **S. Boyd and L. Vandenberghe**, Convex optimization, Cambridge University Press, 2004.
- **H. H. Bauschke and P. L. Combettes**, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer, New York, 2011.
- **F. Bach, R. Jenatton, J. Mairal and G. Obozinski**, Optimization with Sparsity-Inducing Penalties. Foundations and Trends in Machine Learning, 4(1), pages 1?106, 2012. [PDF]
- **P.L. Combettes and J.-C. Pesquet**, Proximal splitting methods in signal processing," in: Fixed-PointAlgorithms for Inverse Problems in Science and Engineering, (H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, Editors), Springer, pp. 185–212, 2011. [PDF]

Optimization
Part II: Basics

Nelly Pustelnik

CNRS, Laboratoire de Physique de l'ENS de Lyon, France

(several slides in this part traced back Tutorial ICASSP 2014 written in collaboration with **Jean-Christophe Pesquet** from Centre de Vision Numérique, CentraleSupelec, University Paris-Saclay, Inria, France.)

Hilbert spaces

A (real) **Hilbert space** $\mathcal{H}$ is a complete real vector space endowed with an inner product $\langle \cdot \mid \cdot \rangle$. The associated norm is

$$(\forall x \in \mathcal{H}) \qquad \|x\| = \sqrt{\langle x \mid x \rangle}.$$

- Particular case: $\mathcal{H} = \mathbb{R}^N$ (Euclidean space with dimension N).

Norm and adjoint

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces.

A linear operator $L\colon \mathcal{H} \to \mathcal{G}$ is **bounded** (or continuous) if

$$\|L\| = \sup_{\|x\|_{\mathcal{H}} \leq 1} \|Lx\|_{\mathcal{G}} < +\infty$$

Norm and adjoint

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces.

A linear operator $L\colon \mathcal{H} \to \mathcal{G}$ is **bounded** (or continuous) if

$$\|L\| = \sup_{\|x\| \leq 1} \|Lx\| < +\infty$$

Norm and adjoint

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces.
A linear operator $L\colon \mathcal{H} \to \mathcal{G}$ is **bounded** (or continuous) if

$$\|L\| = \sup_{\|x\| \leq 1} \|Lx\| < +\infty$$

- In finite dimension, every linear operator is bounded.

Norm and adjoint

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces.
A linear operator $L\colon \mathcal{H} \to \mathcal{G}$ is **bounded** (or continuous) if

$$\|L\| = \sup_{\|x\| \leq 1} \|Lx\| < +\infty$$

- In finite dimension, every linear operator is bounded.

$\mathcal{B}(\mathcal{H}, \mathcal{G})$: Banach space of bounded linear operators from $\mathcal{H}$ to $\mathcal{G}$.

Norm and adjoint

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces.
Let $L \in \mathcal{B}(\mathcal{H}, \mathcal{G})$. Its **adjoint** L^* is the operator in $\mathcal{B}(\mathcal{G}, \mathcal{H})$ defined as

$$(\forall (x, y) \in \mathcal{H} \times \mathcal{G}) \qquad \langle y \mid Lx \rangle_{\mathcal{G}} = \langle L^* y \mid x \rangle_{\mathcal{H}} .$$

Example:

$$\text{If} \qquad L\colon \mathcal{H} \to \mathcal{H}^n \colon x \mapsto (x, \ldots, x)$$

$$\text{then} \qquad L^*\colon \mathcal{H}^n \to \mathcal{H} \colon y = (y_1, \ldots, y_n) \mapsto \sum_{i=1}^{n} y_i$$

Proof: $\langle Lx \mid y \rangle = \langle (x, \ldots, x) \mid (y_1, \ldots, y_n) \rangle = \sum_{i=1}^{n} \langle x \mid y_i \rangle = \left\langle x \;\middle|\; \sum_{i=1}^{n} y_i \right\rangle$

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces.
Let $L \in \mathcal{B}(\mathcal{H}, \mathcal{G})$. Its **adjoint** L^* is the operator in $\mathcal{B}(\mathcal{G}, \mathcal{H})$ defined as

$$(\forall (x, y) \in \mathcal{H} \times \mathcal{G}) \qquad \langle y \mid Lx \rangle = \langle L^* y \mid x \rangle .$$

Example:

If $\quad L : \mathcal{H} \to \mathcal{H}^n : x \mapsto (x, \ldots, x)$

then $\quad L^* : \mathcal{H}^n \to \mathcal{H} : y = (y_1, \ldots, y_n) \mapsto \sum_{i=1}^{n} y_i$

Proof: $\langle Lx \mid y \rangle = \langle (x, \ldots, x) \mid (y_1, \ldots, y_n) \rangle = \sum_{i=1}^{n} \langle x \mid y_i \rangle = \left\langle x \mid \sum_{i=1}^{n} y_i \right\rangle$

Norm and adjoint

- **About L^*:**
  - ⊙ **Compute gradient and proximity operator** operations (Parts III and IV)
  - ⊙ Dual formulation (cf. Part VI)
  - ⊙ Finite dimensions: If $L \in \mathcal{B}(\mathbb{R}^N, \mathbb{R}^K)$ then $L^* = L^\top$.
  - ⊙ Check the correct implementation by using its definition

$$(\forall (x, y) \in \mathbb{R}^N \times \mathbb{R}^K) \qquad \langle Lx \mid y \rangle = \langle x \mid L^* y \rangle$$

- **About $\|L\|$:**
  - ⊙ Required for **gradient-based** algorithms;
  - ⊙ We have $\|L^*\| = \|L\|$;
  - ⊙ **Normalized power method** (or Von Mises iteration) to compute $\|L\|$ when L denotes a diagonalizable matrix.

Norm and adjoint

```
function beta=power_method(H,param)
% Normalized Power Method to estimate ||H||
% Implementation N. Pustelnik
% 23-sept-2020

rhon=1+1e-6;
rhon1(1)=1;
xn  = randn(param.n1,param.n2)';
xn1 = xn;
k=1;
while abs(rhon1(k)-rhon)/rhon1(k) >= 1e-8
    xn  = xn1/norm(xn1,'fro');
    xn1 = H.adj_op((H.dir_op(xn)));
    rhon=rhon1(k);
    k=k+1;
    rhon1(k) = norm(xn1,'fro');
end
beta=sqrt(rhon1(k));
```

Functional analysis: definitions

$$\text{Find} \quad \widehat{x} \in \underset{x \in \mathcal{H}}{\text{Argmin}} \quad f(x)$$

Class of functions $f \in \Gamma_0(\mathcal{H})$:

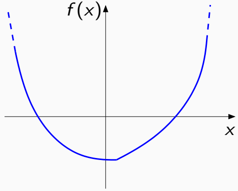
- Proper function
- Lower semi-continuous function
- Convex function

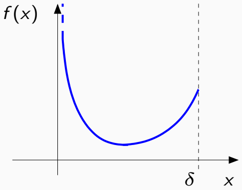
Functional analysis: definitions

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$ where $\mathcal{H}$ is a Hilbert space.

- The **domain** of f is $\operatorname{dom} f = \{x \in \mathcal{H} \,|\, f(x) < +\infty\}$.
- The function f is **proper** if $\operatorname{dom} f \neq \varnothing$.

Domains of the functions ?



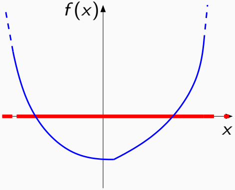


Functional analysis: definitions

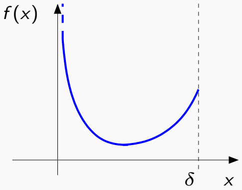
Let $f : \mathcal{H} \to \,]-\infty, +\infty]$ where $\mathcal{H}$ is a Hilbert space.

- The **domain** of f is $\operatorname{dom} f = \{x \in \mathcal{H} \,|\, f(x) < +\infty\}$.
- The function f is **proper** if $\operatorname{dom} f \neq \varnothing$.

Domains of the functions ?



$\operatorname{dom} f = \mathbb{R}$
(proper)

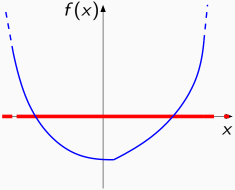


Functional analysis: definitions

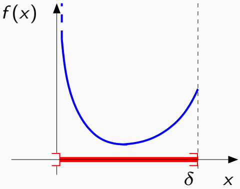
Let $f : \mathcal{H} \to \,]-\infty, +\infty]$ where $\mathcal{H}$ is a Hilbert space.

- The **domain** of f is $\operatorname{dom} f = \{x \in \mathcal{H} \mid f(x) < +\infty\}$.
- The function f is **proper** if $\operatorname{dom} f \neq \varnothing$.

Domains of the functions ?



$\operatorname{dom} f = \mathbb{R}$
(proper)



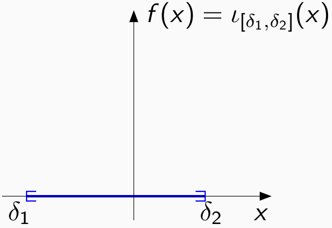
$\operatorname{dom} f =]0, \delta]$
(proper)

Functional analysis: definitions

Let $C \subset \mathcal{H}$.
The **indicator function of C** is

$$(\forall x \in \mathcal{H}) \qquad \iota_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{otherwise.} \end{cases}$$

Example : $C = [\delta_1, \delta_2]$



Epigraph

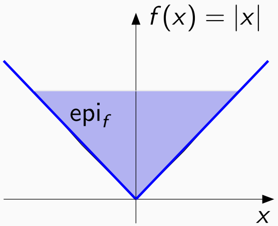
Let $f : \mathcal{H} \to \left]-\infty, +\infty\right]$. The **epigraph** of f is

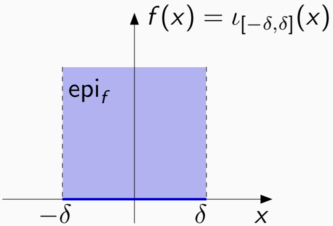
$$\operatorname{epi} f = \left\{ (x, \zeta) \in \operatorname{dom} f \times \mathbb{R} \;\middle|\; f(x) \leq \zeta \right\}$$

Epigraph

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$. The **epigraph** of f is

$$\operatorname{epi} f = \left\{(x, \zeta) \in \operatorname{dom} f \times \mathbb{R} \;\middle|\; f(x) \leq \zeta\right\}$$

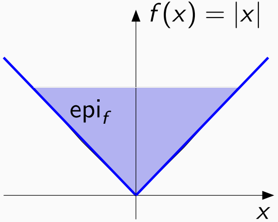


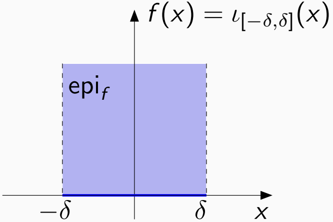


Epigraph

Let $f : \mathcal{H} \to\ ]-\infty, +\infty]$. The **epigraph** of f is

$$\text{epi}\, f = \left\{(x, \zeta) \in \operatorname{dom} f \times \mathbb{R} \;\middle|\; f(x) \leq \zeta\right\}$$





- Examples:
  - ⊙ Astrophysics: Epigraphical constraint on **Stokes parameters** $x = (I, Q, U)$: $I_n \geq \sqrt{Q_n^2 + U_n^2}$
  - ⊙ **Projection onto ℓ_1-ball**: $\sum_n |x_n| \leq \eta \Leftrightarrow \begin{cases} |x_n| \leq \zeta_n \\ \sum_n \zeta_n \leq \eta \end{cases}$

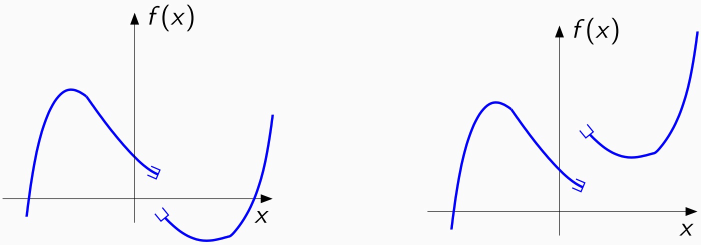
Lower semi-continuity

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$.
f is a **lower semi-continuous** function on $\mathcal{H}$ if and only if $\operatorname{epi} f$ is closed

Lower semi-continuity

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$.
f is a **lower semi-continuous** function on $\mathcal{H}$ if and only if $\operatorname{epi} f$ is closed

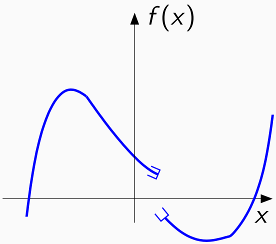
- l.s.c. functions ?

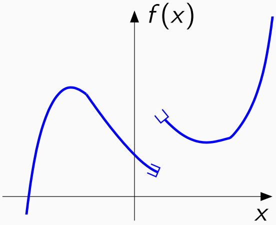


Lower semi-continuity

Let $f : \mathcal{H} \to]-\infty, +\infty]$.
f is a **lower semi-continuous** function on $\mathcal{H}$ if and only if $\operatorname{epi} f$ is closed

- l.s.c. functions ?

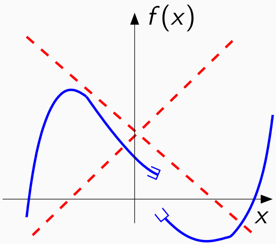


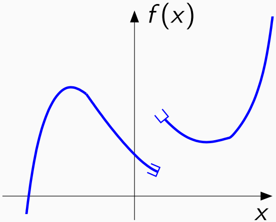


Lower semi-continuity

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$.
f is a **lower semi-continuous** function on $\mathcal{H}$ if and only if $\operatorname{epi} f$ is closed

- l.s.c. functions ?

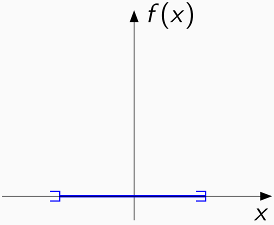


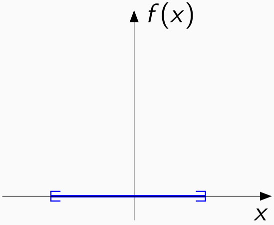


Lower semi-continuity

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$.
f is a **lower semi-continuous** function on $\mathcal{H}$ if and only if $\operatorname{epi} f$ is closed

- l.s.c. functions ?

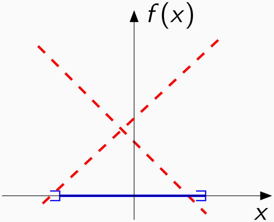


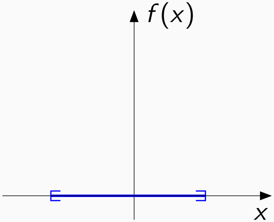


Lower semi-continuity

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$.
f is a **lower semi-continuous** function on $\mathcal{H}$ if and only if epi f is closed

- l.s.c. functions ?

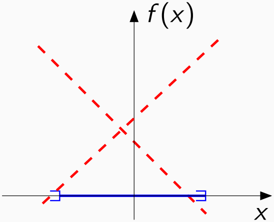


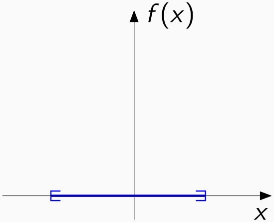


Lower semi-continuity

Let $f : \mathcal{H} \to]-\infty, +\infty]$.
f is a **lower semi-continuous** function on $\mathcal{H}$ if and only if $\operatorname{epi} f$ is closed

- l.s.c. functions ?





Lower semi-continuity

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$.
f is a **lower semi-continuous** function on $\mathcal{H}$ if and only if epi f is closed

- Examples:
  - ⊙ **Do not allow for strict constraints** e.g. $Ax < b$ or $x > 0$;
  - ⊙ Allow for inequality or equality constraints e.g. $Ax = b$, $Ax \leq b$ or $x > 0$;

Lower semi-continuity

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$.
f is a **lower semi-continuous** function on $\mathcal{H}$ if and only if $\operatorname{epi} f$ is closed

- Examples:
  - ⊙ **Do not allow for strict constraints** e.g. $Ax < b$ or $x > 0$;
  - ⊙ Allow for inequality or equality constraints e.g. $Ax = b$, $Ax \leq b$ or $x > 0$;

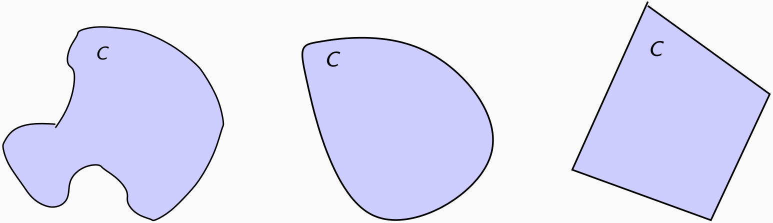
- Properties:
  - ⊙ Every continuous function on $\mathcal{H}$ is l.s.c.
  - ⊙ **Every finite sum of l.s.c. functions is l.s.c.**
  - ⊙ Let $(f_i)_{i \in I}$ be a family of l.s.c functions. Then, $\sup_{i \in I} f_i$ is l.s.c.

Convex set

$C \subset \mathcal{H}$ is a **convex set** if

$$(\forall (x, y) \in C^2)(\forall \alpha \in]0, 1[) \qquad \alpha x + (1 - \alpha) y \in C$$

Convex sets ?

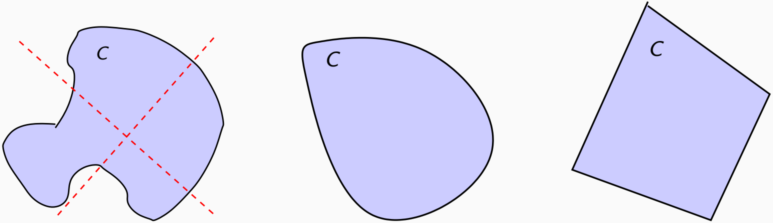


Convex set

$C \subset \mathcal{H}$ is a **convex set** if

$$(\forall (x, y) \in C^2)(\forall \alpha \in]0, 1[) \qquad \alpha x + (1 - \alpha) y \in C$$

Convex sets ?



Convex function: definitions

$f : \mathcal{H} \to \,]-\infty, +\infty]$ is a **convex function** if

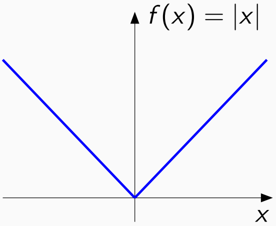
$$(\forall (x, y) \in \mathcal{H}^2)(\forall \alpha \in]0, 1[) \qquad f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

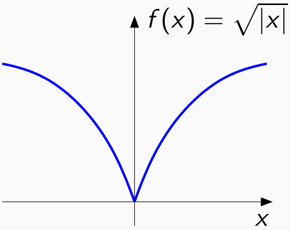
Convex function: definitions

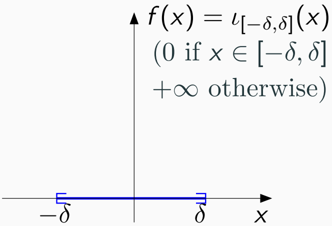
$f : \mathcal{H} \to \,]-\infty, +\infty]$ is a **convex function** if

$$(\forall (x,y) \in \mathcal{H}^2)(\forall \alpha \in]0,1[) \qquad f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

Convex functions ?





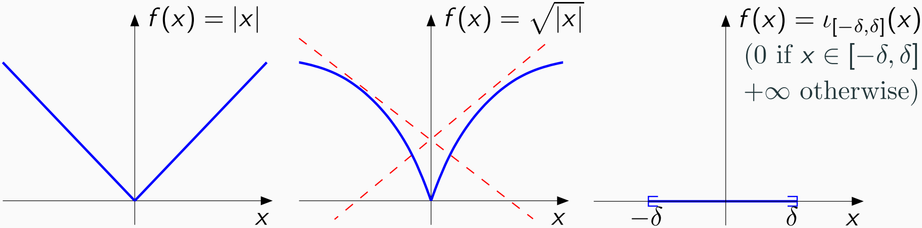


Convex function: definitions

$f : \mathcal{H} \to \,]-\infty, +\infty]$ is a **convex function** if

$$(\forall (x, y) \in \mathcal{H}^2)(\forall \alpha \in]0, 1[) \qquad f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y)$$

Convex functions ?

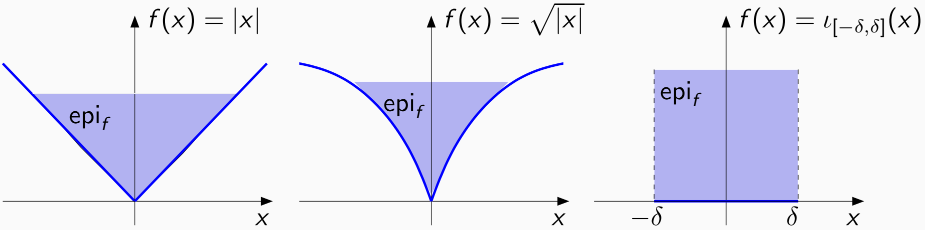


Convex functions: definition

$f : \mathcal{H} \to \,]-\infty, +\infty]$ is convex $\Leftrightarrow$ its epigraph is convex.

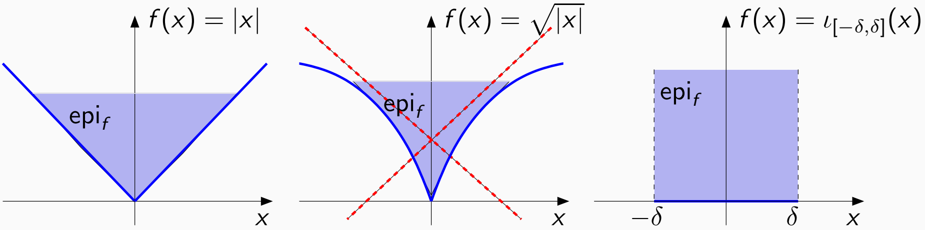
Convex functions: definition

$f : \mathcal{H} \to \,]-\infty, +\infty]$ is convex $\Leftrightarrow$ its epigraph is convex.



Convex functions: definition

$f : \mathcal{H} \to]-\infty, +\infty]$ is convex $\Leftrightarrow$ its epigraph is convex.



Convex functions: definition

$f : \mathcal{H} \to \,]-\infty, +\infty]$ is convex $\Leftrightarrow$ its epigraph is convex.

- Properties :
  - ⊙ Composition of an increasing convex funct. and a convex funct. is convex.
  - ⊙ If $f : \mathcal{H} \to \,]-\infty, +\infty]$ is convex, then $\operatorname{dom} f$ is convex.
  - ⊙ $f : \mathcal{H} \to [-\infty, +\infty[$ is concave if $-f$ is convex.
  - ⊙ Every finite **sum of convex functions is convex**.
  - ⊙ Let $(f_i)_{i \in I}$ be a family of convex functions. Then, $\sup_{i \in I} f_i$ is convex.

- $\Gamma_0(\mathcal{H})$: class of convex, l.s.c., and proper functions from $\mathcal{H}$ to $]-\infty, +\infty]$.

- $\iota_C \in \Gamma_0(\mathcal{H}) \Leftrightarrow C$ is a nonempty closed convex set.

Strictly convex functions

Let $\mathcal{H}$ be a Hilbert space. Let $f\colon \mathcal{H} \to\ ]-\infty, +\infty]$.
f is **strictly convex** if

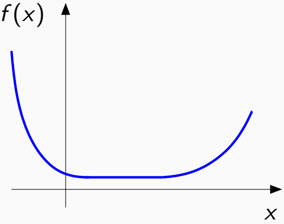
$$(\forall x \in \operatorname{dom} f)(\forall y \in \operatorname{dom} f)(\forall \alpha \in]0,1[)$$
$$x \neq y \quad \Rightarrow \quad f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y).$$

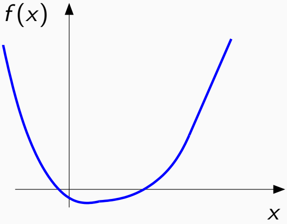
Strictly convex functions

Let $\mathcal{H}$ be a Hilbert space. Let $f\colon \mathcal{H} \to \,]-\infty, +\infty]$.
f is **strictly convex** if

$$(\forall x \in \operatorname{dom} f)(\forall y \in \operatorname{dom} f)(\forall \alpha \in]0,1[)$$
$$x \neq y \quad \Rightarrow \quad f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y).$$

- Strictly convex functions ?



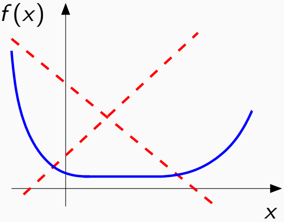


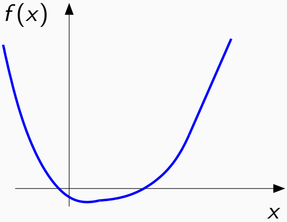
Strictly convex functions

Let $\mathcal{H}$ be a Hilbert space. Let $f\colon \mathcal{H} \to\ ]-\infty, +\infty]$.
f is **strictly convex** if

$$(\forall x \in \operatorname{dom} f)(\forall y \in \operatorname{dom} f)(\forall \alpha \in]0,1[)$$
$$x \neq y \quad \Rightarrow \quad f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y).$$

- Strictly convex functions ?





Functional analysis: minimizers

$$\text{Find} \quad \widehat{x} \in \underset{x \in C}{\operatorname{Argmin}} \quad f(x)$$

- Class of functions $f \in \Gamma_0(\mathcal{H})$:

- **Minimizers**
  - ⊙ Local versus global minimizers
  - ⊙ Coercivity and existence
  - ⊙ Convex function

Minimizers

Let C be a nonempty set of a Hilbert space $\mathcal{H}$.
Let $f : C \to \,]-\infty, +\infty]$ be a proper function and let $\widehat{x} \in C$.

- $\widehat{x} \in \operatorname{dom} f$ is a **local minimizer** of f if there exists an open neigborhood O of $\widehat{x}$ such that
$$(\forall x \in O \cap C) \qquad f(\widehat{x}) \leq f(x).$$
- $\widehat{x}$ is a **(global) minimizer** of f if
$$(\forall x \in C) \qquad f(\widehat{x}) \leq f(x).$$

Minimizers

Let C be a nonempty set of a Hilbert space $\mathcal{H}$.
Let $f : C \to]-\infty, +\infty]$ be a proper function and let $\widehat{x} \in C$.

- $\widehat{x}$ is a **strict local minimizer** of f if there exists an open neigborhood O of $\widehat{x}$ such that
$$(\forall x \in (O \cap C) \setminus \{\widehat{x}\}) \qquad f(\widehat{x}) < f(x).$$
- $\widehat{x}$ is a **strict (global) minimizer** of f if
$$(\forall x \in C \setminus \{\widehat{x}\}) \qquad f(\widehat{x}) < f(x).$$

Minimizers of a convex function

Theorem: Let $\mathcal{H}$ be a Hilbert space. Let $f : \mathcal{H} \to]-\infty, +\infty]$ be a **proper convex** function such that $\mu = \inf f > -\infty$.

- $\{x \in \mathcal{H} \mid f(x) = \mu\}$ is convex.
- Every local minimizer of f is a global minimizer.
- If f is strictly convex, then there exists at most one minimizer.

Minimizers of a convex function

Theorem: Let $\mathcal{H}$ be a Hilbert space. Let $f\colon \mathcal{H} \to \left]-\infty, +\infty\right]$ be a **proper convex** function such that $\mu = \inf f > -\infty$.

- $\{x \in \mathcal{H} \mid f(x) = \mu\}$ is convex.
- Every local minimizer of f is a global minimizer.
- If f is strictly convex, then there exists at most one minimizer.

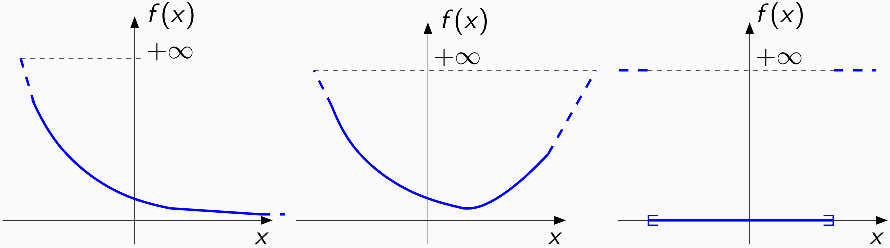
Existence of a minimizer

Let $\mathcal{H}$ be a Hilbert space. Let $f\colon \mathcal{H} \to \,]-\infty, +\infty]$.
f is **coercive** if $\lim_{\|x\| \to +\infty} f(x) = +\infty$.

Existence of a minimizer

Let $\mathcal{H}$ be a Hilbert space. Let $f\colon \mathcal{H} \to \,]-\infty, +\infty]$.
f is **coercive** if $\lim_{\|x\| \to +\infty} f(x) = +\infty$.

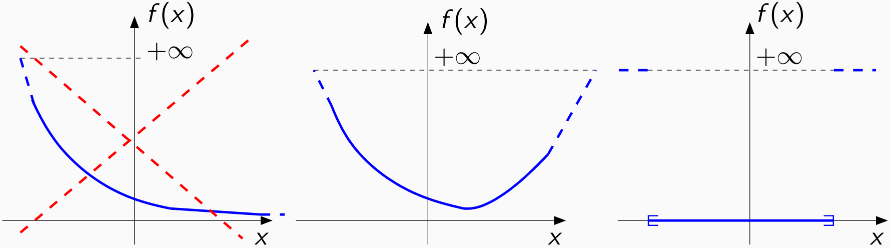
Coercive functions ?



Existence of a minimizer

Let $\mathcal{H}$ be a Hilbert space. Let $f\colon \mathcal{H} \to \,]-\infty, +\infty]$.
f is **coercive** if $\lim_{\|x\|\to+\infty} f(x) = +\infty$.

Coercive functions ?



Existence and uniqueness of a minimizer

Theorem: Let $\mathcal{H}$ be a Hilbert space and C a **closed convex** subset of $\mathcal{H}$. Let $f \in \Gamma_0(\mathcal{H})$ such that $\mathrm{dom}\, f \cap C \neq \varnothing$.
If f **is coercive** or C **is bounded**, then there exists $\widehat{x} \in C$ such that

$$f(\widehat{x}) = \inf_{x \in C} f(x).$$

If, moreover, f is strictly convex, this minimizer $\widehat{x}$ is unique.

Functional analysis: minimizers

$$\text{Find} \quad \widehat{x} \in \underset{x \in C}{\mathrm{Argmin}} \quad f(x)$$

- Class of functions $f \in \Gamma_0(\mathcal{H})$:
- Minimizers
- **Differentiability and optimality condition**

Differentiable functions

Let $\mathcal{H}$ be a Hilbert space and let $f\colon \mathcal{H} \to \,]-\infty, +\infty]$ be a proper function.
f is **Gâteaux differentiable** at $x \in \operatorname{dom} f$ if there exists $\nabla f(x) \in \mathcal{H}$ such that

$$(\forall y \in \mathcal{H}) \qquad \langle \nabla f(x) \mid y \rangle = \lim_{\substack{\alpha \to 0 \\ \alpha \neq 0}} \frac{f(x + \alpha y) - f(x)}{\alpha}.$$

- $\nabla f(x) \in \mathcal{H}$ the Riesz-Fréchet representant
- Example: Let $x \in \mathbb{R}^N$, $z \in \mathbb{R}^K$ and $A \in \mathbb{R}^{K \times N}$ and $f(x) = \frac{1}{2}\|Ax - z\|^2$, then

$$\nabla f(x) = A^*(Ax - z)$$

Optimality condition

1st order necessary and sufficient condition (P. Fermat)

Let $f \in \Gamma_0(\mathbb{R}^N)$ be continuously differentiable function on $\mathbb{R}^N$.

$$\widehat{x} \in \operatorname{Argmin}_{x \in \mathbb{R}^N} f(x) \quad \Leftrightarrow \quad \nabla f(\widehat{x}) = 0$$

- More details about optimality conditions here :
  [Jean-Charles Gilbert course]
  [Nocedal-Wright, 1999]

- Limitations :
  - ⊙ Lead to a N equations - N unknown problem.
  - ⊙ Closed form expression for only few cases.
  - ⊙ If no closed form expression exists, an iterative procedure is required.

Optimality condition

- Example: **Solving mean squares**

Find $\quad \widehat{x} = \mathrm{Argmin}_{x \in \mathbb{R}^N} \|Ax - y\|_2^2 \qquad$ with $\qquad \begin{cases} A \in \mathbb{R}^{N \times N} & \text{full rank} \\ y \in \mathbb{R}^M \end{cases}$

→ Optimality condition:

$$\nabla f(\widehat{x}) = 0 \quad \Leftrightarrow \quad A^\top (A\widehat{x} - y) = 0$$

$$\boxed{\widehat{x} = (A^\top A)^{-1}(A^\top y)}$$

→ **Closed form expression** but sometimes difficult to invert $A^\top A$.

Optimality condition

- Example: **Logistic based criterion:**

$$\text{Find} \quad \widehat{x} \in \mathrm{Argmin}_{x\in\mathbb{R}} \log\big(1+\exp(-yx)\big) \qquad \text{with} \qquad y \in \mathbb{R}$$

→ Optimality condition:

$$\nabla f(\widehat{x}) = 0 \quad \Leftrightarrow \quad \boxed{\frac{-y\exp(-y\widehat{x})}{1+\exp(-y\widehat{x})} = 0}$$

→ **No closed form expression**. An iterative procedure is required.

Iterative scheme

Problem: Let $f \in \Gamma_0(\mathbb{R}^N)$, find $\quad \widehat{x} \in \underset{x \in \mathbb{R}^N}{\mathrm{Argmin}}\, f(x)$.

- If f is α-Lipschitz differentiable with $\alpha > 0$, the (explicit) **gradient method**:

$$(\forall n \in \mathbb{N}) \qquad x_{n+1} = x_n - \gamma_n \nabla f(x_n)$$

$\rightarrow$ Convergence insured when $0 < \inf_{n \in \mathbb{N}} \gamma_n$ et $\sup_{n \in \mathbb{N}} \gamma_n < 2\alpha^{-1}$.

Iterative scheme

Problem: Let $f \in \Gamma_0(\mathbb{R}^N)$, find $\quad \widehat{x} \in \operatorname{Argmin}_{x \in \mathbb{R}^N} f(x)$.

- If f is α-Lipschitz differentiable with $\alpha > 0$, the (explicit) **gradient method**:

$$(\forall n \in \mathbb{N}) \qquad x_{n+1} = x_n - \gamma_n \nabla f(x_n)$$

$\rightarrow$ Convergence insured when $0 < \inf_{n \in \mathbb{N}} \gamma_n$ et $\sup_{n \in \mathbb{N}} \gamma_n < 2\alpha^{-1}$.

- If f nonsmooth, the (explicit) **subgradient method**:

$$(\forall n \in \mathbb{N}) \qquad x_{n+1} = x_n - \gamma_n u_n \quad \text{with} \quad u_n \in \partial f(x_n)$$

$\rightarrow$ Convergence insured when $\gamma_n \in]0, +\infty[$ such that $\sum_{n=0}^{+\infty} \gamma_n^2 < +\infty$ and $\sum_{n=0}^{+\infty} \gamma_n = +\infty$. [Shor, 1979].

Iterative scheme

Problem: Let $f \in \Gamma_0(\mathbb{R}^N)$, find $\quad \widehat{x} \in \underset{x\in\mathbb{R}^N}{\mathrm{Argmin}}\, f(x)$.

- If f is α-Lipschitz differentiable with $\alpha > 0$, the (explicit) **gradient method**:

$$(\forall n \in \mathbb{N}) \qquad x_{n+1} = x_n - \gamma_n \nabla f(x_n)$$

$\rightarrow$ Convergence insured when $0 < \inf_{n\in\mathbb{N}} \gamma_n$ et $\sup_{n\in\mathbb{N}} \gamma_n < 2\alpha^{-1}$.

- If f nonsmooth, the (explicit) **subgradient method**:

$$(\forall n \in \mathbb{N}) \qquad x_{n+1} = x_n - \gamma_n u_n \quad \text{with} \quad u_n \in \partial f(x_n)$$

$\rightarrow$ Convergence insured when $\gamma_n \in]0, +\infty[$ such that $\sum_{n=0}^{+\infty} \gamma_n^2 < +\infty$ and $\sum_{n=0}^{+\infty} \gamma_n = +\infty$. [Shor, 1979].

- If f nonsmooth, the **implicit subgradient method** is

$$(\forall n \in \mathbb{N}) \qquad x_{n+1} = x_n - \gamma_n u_n \quad \text{with} \quad u_n \in \partial f(x_{n+1})$$

$\rightarrow$ Convergence insured when $\sum_{n=0}^{+\infty} \gamma_n = +\infty$.

Iterative scheme

Problem: Let $f \in \Gamma_0(\mathbb{R}^N)$, find $\quad \widehat{x} \in \underset{x \in \mathbb{R}^N}{\operatorname{Argmin}} f(x)$.

- If f is α-Lipschitz differentiable with $\alpha > 0$, the (explicit) **gradient method**:

$$(\forall n \in \mathbb{N}) \qquad x_{n+1} = x_n - \gamma_n \nabla f(x_n)$$

$\rightarrow$ Convergence insured when $0 < \inf_{n \in \mathbb{N}} \gamma_n$ et $\sup_{n \in \mathbb{N}} \gamma_n < 2\alpha^{-1}$.

- If f nonsmooth, the (explicit) **subgradient method**:

$$(\forall n \in \mathbb{N}) \qquad x_{n+1} = x_n - \gamma_n u_n \quad \text{with} \quad u_n \in \partial f(x_n)$$

$\rightarrow$ Convergence insured when $\gamma_n \in]0, +\infty[$ such that $\sum_{n=0}^{+\infty} \gamma_n^2 < +\infty$ and $\sum_{n=0}^{+\infty} \gamma_n = +\infty$. [Shor, 1979].

- If f nonsmooth, the **implicit subgradient method** is

$$(\forall n \in \mathbb{N}) \qquad x_{n+1} = x_n - \gamma_n u_n \quad \text{with} \quad u_n \in \partial f(x_{n+1})$$

$\rightarrow$ Convergence insured when $\sum_{n=0}^{+\infty} \gamma_n = +\infty \Rightarrow$ **Proximity operator**.

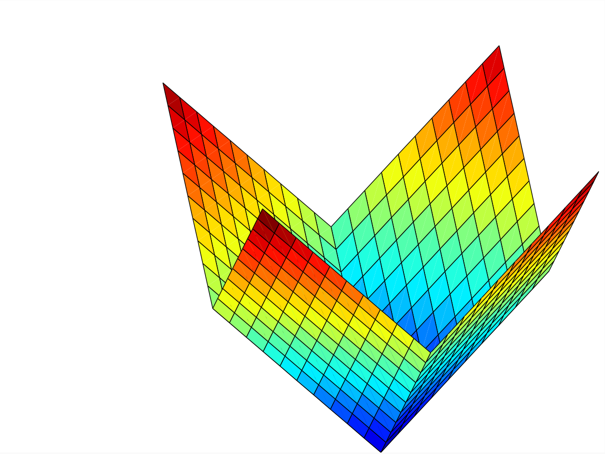
Optimization
Part III: Subdifferential

Nelly Pustelnik

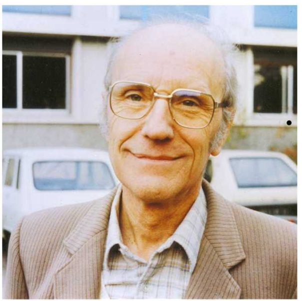
CNRS, Laboratoire de Physique de l'ENS de Lyon, France

(several slides in this part traced back Tutorial ICASSP 2014 written in collaboration with **Jean-Christophe Pesquet** from Centre de Vision Numérique, CentraleSupelec, University Paris-Saclay, Inria, France.)

Non-smooth convex optimization



A pioneer



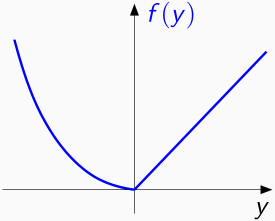
Jean-Jacques Moreau
(1923–2014)

Subdifferential of function: definition

The (Moreau) **subdifferential of f**, denoted by ∂f,

Subdifferential of function: definition

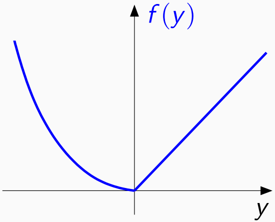
Let $f : \mathcal{H} \to \,]-\infty, +\infty]$ be a proper function.
The (Moreau) **subdifferential of f**, denoted by ∂f,

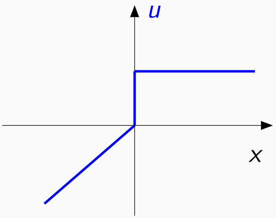


Subdifferential of function: definition

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$ be a proper function.
The (Moreau) **subdifferential of f**, denoted by ∂f, is such that

$$\partial f : \mathcal{H} \to 2^{\mathcal{H}}$$
$$x \to \{u \in \mathcal{H} \mid (\forall y \in \mathcal{H})\ \langle y - x | u \rangle + f(x) \leqslant f(y)\}$$

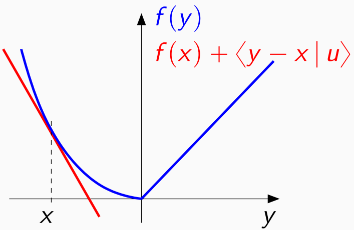


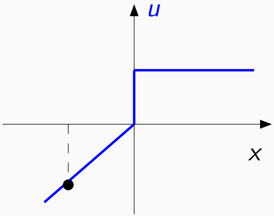


Subdifferential of function: definition

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$ be a proper function.
The (Moreau) **subdifferential of f**, denoted by ∂f, is such that

$$\partial f : \mathcal{H} \to 2^{\mathcal{H}}$$
$$x \to \{u \in \mathcal{H} \mid (\forall y \in \mathcal{H})\ \langle y - x | u\rangle + f(x) \leqslant f(y)\}$$

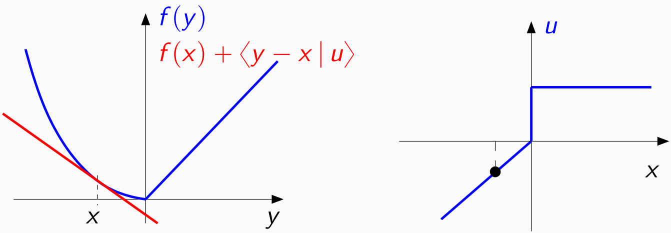




Subdifferential of function: definition

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$ be a proper function.
The (Moreau) **subdifferential of f**, denoted by ∂f, is such that

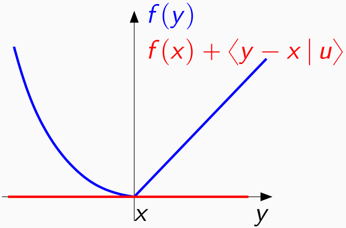
$$\begin{aligned} \partial f : \mathcal{H} &\to 2^{\mathcal{H}} \\ x &\to \{u \in \mathcal{H} \mid (\forall y \in \mathcal{H})\ \langle y - x | u\rangle + f(x) \leqslant f(y)\} \end{aligned}$$

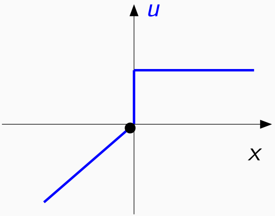


Subdifferential of function: definition

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$ be a proper function.
The (Moreau) **subdifferential of f**, denoted by ∂f, is such that

$$\begin{aligned} \partial f : \mathcal{H} &\to 2^{\mathcal{H}} \\ x &\to \{u \in \mathcal{H} \mid (\forall y \in \mathcal{H})\ \langle y - x | u \rangle + f(x) \leqslant f(y)\} \end{aligned}$$

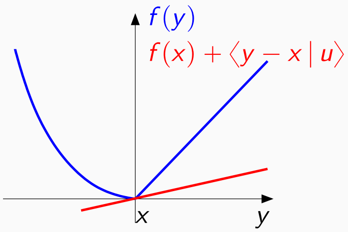


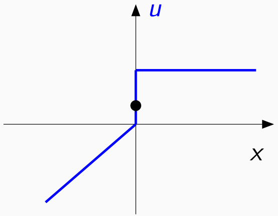


Subdifferential of function: definition

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$ be a proper function.
The (Moreau) **subdifferential of f**, denoted by ∂f, is such that

$$\begin{aligned} \partial f : \mathcal{H} &\to 2^{\mathcal{H}} \\ x &\to \{u \in \mathcal{H} \mid (\forall y \in \mathcal{H})\ \langle y - x | u \rangle + f(x) \leqslant f(y)\} \end{aligned}$$

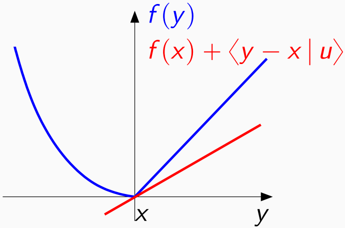


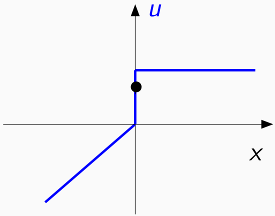


Subdifferential of function: definition

Let $f : \mathcal{H} \to]-\infty, +\infty]$ be a proper function.
The (Moreau) **subdifferential of f**, denoted by ∂f, is such that

$$\begin{aligned} \partial f : \mathcal{H} &\to 2^{\mathcal{H}} \\ x &\to \{u \in \mathcal{H} \mid (\forall y \in \mathcal{H})\ \langle y - x | u \rangle + f(x) \leqslant f(y)\} \end{aligned}$$

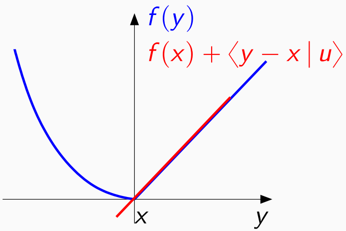


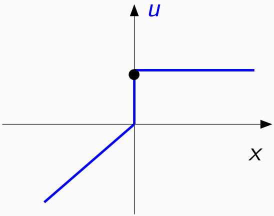


Subdifferential of function: definition

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$ be a proper function.
The (Moreau) **subdifferential of f**, denoted by ∂f, is such that

$$\begin{aligned} \partial f : \mathcal{H} &\to 2^{\mathcal{H}} \\ x &\to \{u \in \mathcal{H} \mid (\forall y \in \mathcal{H})\ \langle y - x | u \rangle + f(x) \leqslant f(y)\} \end{aligned}$$

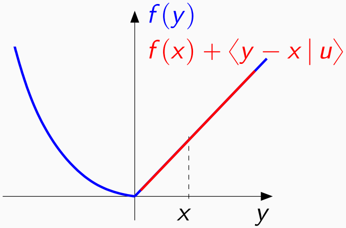


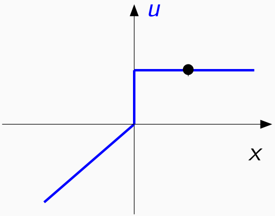


Subdifferential of function: definition

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$ be a proper function.
The (Moreau) **subdifferential of f**, denoted by ∂f, is such that

$$\begin{aligned} \partial f : \mathcal{H} &\to 2^{\mathcal{H}} \\ x &\to \{u \in \mathcal{H} \mid (\forall y \in \mathcal{H})\ \langle y - x | u \rangle + f(x) \leqslant f(y)\} \end{aligned}$$

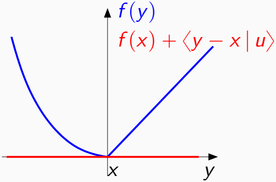


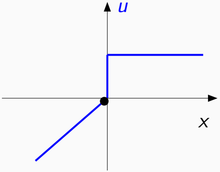


Subdifferential of a function: properties

Let $f : \mathcal{H} \to]-\infty, +\infty]$ be a proper function.
The (Moreau) subdifferential of f, denoted by ∂f, is such that

$$\begin{aligned} \partial f : \mathcal{H} &\to 2^{\mathcal{H}} \\ x &\to \{u \in \mathcal{H} \,|\, (\forall y \in \mathcal{H})\ \langle y - x | u\rangle + f(x) \leqslant f(y)\} \end{aligned}$$





Fermat's rule: $0 \in \partial f(\widehat{x}) \Leftrightarrow \widehat{x} \in \underset{x}{\text{Argmin}}\ f(x)$

Subdifferential of a function: properties

Let $f : \mathcal{H} \to \,]-\infty, +\infty]$ be a proper function.
The (Moreau) subdifferential of f, denoted ∂f, is such that

$$\begin{aligned} \partial f : \mathcal{H} &\to 2^{\mathcal{H}} \\ x &\to \{u \in \mathcal{H} \mid (\forall y \in \mathcal{H})\ \langle y - x | u \rangle + f(x) \leqslant f(y)\} \end{aligned}$$

- $u \in \partial f(x)$ is a subgradient of f at x.

Subdifferential of a convex function: properties

If $f\colon \mathcal{H} \to \left]-\infty, +\infty\right]$ is convex and it is Gâteaux differentiable at x, then

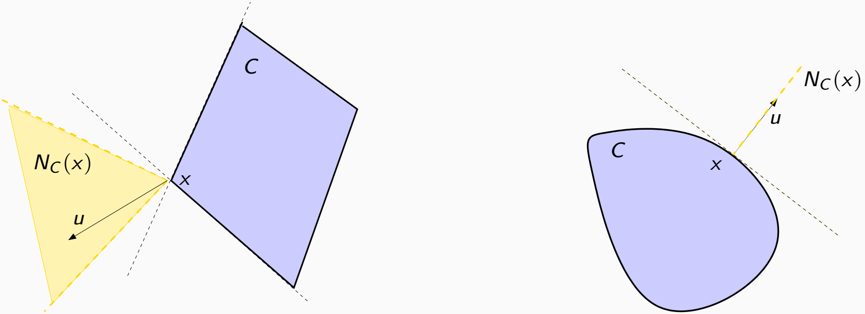
$$\partial f(x) = \{\nabla f(x)\}$$

Subdifferential of a convex function: example

Let C be a nonempty subset of $\mathcal{H}$.

For every $x \in \mathcal{H}$, $\partial\iota_C(x)$ is the normal cone to C at x defined by

$$N_C(x) = \begin{cases} \{u \in \mathcal{H} \mid (\forall y \in C) \;\; \langle u \mid y - x\rangle \leqslant 0\} & \text{if } x \in C \\ \varnothing & \text{otherwise.} \end{cases}$$



Subdifferential calculus

Let $\mathcal{H}$ and $\mathcal{G}$ be two real Hilbert spaces.

- Let $f\colon \mathcal{H} \to \,]-\infty,+\infty]$ be proper, then for every $\lambda \in \,]0,+\infty[$ $\partial(\lambda f) = \lambda \partial f$.
- Let $f\colon \mathcal{H} \to \,]-\infty,+\infty]$, $g\colon \mathcal{G} \to \,]-\infty,+\infty]$, and $L \in \mathcal{B}(\mathcal{H},\mathcal{G})$.
  If $\mathit{dom}\, g \cap L(\mathit{dom}\, f) \neq \varnothing$, then

$$(\forall x \in \mathcal{H}) \qquad \partial f(x) + L^* \partial g(Lx) \subset \partial(f + g \circ L)(x).$$

Subdifferential calculus

Let $\mathcal{H}$ and $\mathcal{G}$ be two real Hilbert spaces.
Let $f \in \Gamma_0(\mathcal{H})$, $g \in \Gamma_0(\mathcal{G})$, and $L \in \mathcal{B}(\mathcal{H}, \mathcal{G})$.
If $int\,(dom\,g) \cap L(dom\,f) \neq \varnothing$ or $dom\,g \cap int\,\big(L(dom\,f)\big) \neq \varnothing$, then

$$\partial f + L^* \partial g\, L = \partial(f + g \circ L).$$

Particular case:

- If $f \in \Gamma_0(\mathcal{H})$, $g \in \Gamma_0(\mathcal{H})$, and $dom\,g = \mathcal{H}$ (or $dom\,f = \mathcal{H}$), then $\partial f + \partial g = \partial(f + g)$.

- If $g \in \Gamma_0(\mathcal{G})$, $L \in \mathcal{B}(\mathcal{G}, \mathcal{H})$, and $int\,(dom\,g) \cap ran\,L \neq \varnothing$, then $L^* \partial g\, L = \partial(g \circ L)$.

Subdifferential calculus

Let $(\mathcal{H})_{i\in I}$ where $I \subset \mathbb{N}$ be Hilbert spaces and let $\mathcal{H} = \bigoplus_{i\in I} \mathcal{H}_i$.
For every $i \in I$, let $f_i \colon \mathcal{H}_i \to \,]-\infty, +\infty]$ be a proper function. Let

$$f \colon \mathcal{H} \to \,]-\infty, +\infty] : x = (x_i)_{i\in I} \mapsto \sum_{i\in I} f_i(x_i)$$

Then,

$$\big(\forall x = (x_i)_{i\in I} \in \mathcal{H}\big) \qquad \partial f(x) = \bigtimes_{i\in I} \partial f_i(x_i).$$

Subdifferential calculus

Let $(\mathcal{H})_{i\in I}$ where $I \subset \mathbb{N}$ be Hilbert spaces and let $\mathcal{H} = \bigoplus_{i\in I} \mathcal{H}_i$.
For every $i \in I$, let $f_i \colon \mathcal{H}_i \to \left]-\infty, +\infty\right]$ be a proper function. Let

$$f \colon \mathcal{H} \to \left]-\infty, +\infty\right] : x = (x_i)_{i\in I} \mapsto \sum_{i\in I} f_i(x_i)$$

Then,

$$\big(\forall x = (x_i)_{i\in I} \in \mathcal{H}\big) \qquad \partial f(x) = \bigtimes_{i\in I} \partial f_i(x_i).$$

Proof: Let $x = (x_i)_{i\in I} \in \mathcal{H}$. We have

$$\begin{aligned}
& t = (t_i)_{i\in I} \in \bigtimes_{i\in I} \partial f_i(x_i) \\
\Leftrightarrow\quad & (\forall i \in I)(\forall y_i \in \mathcal{H}_i) \;\; f_i(y_i) \geqslant f_i(x_i) + \langle t_i \mid y_i - x_i \rangle \\
\Rightarrow\quad & \big(\forall y = (y_i)_{i\in I} \in \mathcal{H}\big) \;\; \sum_{i\in I} f_i(y_i) \geqslant \sum_{i\in I} f_i(x_i) + \sum_{i\in I} \langle t_i \mid y_i - x_i \rangle \\
\Leftrightarrow\quad & \big(\forall y \in \mathcal{H}\big) \;\; f(y) \geqslant f(x) + \langle t \mid y - x \rangle.
\end{aligned}$$

Subdifferential calculus

Let $(\mathcal{H})_{i\in I}$ where $I \subset \mathbb{N}$ be Hilbert spaces and let $\mathcal{H} = \bigoplus_{i\in I} \mathcal{H}_i$.
For every $i \in I$, let $f_i \colon \mathcal{H}_i \to \left]-\infty, +\infty\right]$ be a proper function. Let

$$f \colon \mathcal{H} \to \left]-\infty, +\infty\right] : x = (x_i)_{i\in I} \mapsto \sum_{i\in I} f_i(x_i)$$

Then,

$$\big(\forall x = (x_i)_{i\in I} \in \mathcal{H}\big) \qquad \partial f(x) = \bigtimes_{i\in I} \partial f_i(x_i).$$

Proof: Conversely,

$$\begin{aligned} & t = (t_i)_{i\in I} \in \partial f(x) \\ \Leftrightarrow \quad & \big(\forall y = (y_i)_{i\in I} \in \mathcal{H}\big) \;\; \sum_{i\in I} f_i(y_i) \geqslant \sum_{i\in I} f_i(x_i) + \sum_{i\in I} \langle t_i \mid y_i - x_i \rangle . \end{aligned}$$

Let $j \in I$. By setting $(\forall i \in I \backslash \{j\})$ $y_i = x_i \in \mathit{dom}\, f_i$, we get

$$(\forall y_j \in \mathcal{H}_j) \;\; f_j(y_j) \geqslant f_j(x_j) + \langle t_j \mid y_j - x_j \rangle .$$

L1 norm

→ **ℓ_1−norm**

$$f : \mathbb{R}^N \to \mathbb{R} : (x_i)_{1\leqslant i\leqslant N} \mapsto \sum_{i=1}^{N} |x_i|$$

Then

$$\partial|\cdot| : \zeta \mapsto \begin{cases} -1 & \text{if } \zeta < 0; \\ [-1,1] & \text{if } \zeta = 0, \\ 1 & \text{if } \zeta > 0; \end{cases}$$

Huber function

→ **Smooth approximation of the ℓ_1−norm** parametrized by $\mu > 0$. [**Combettes-Glaudin,2019**]

$$f : \mathbb{R}^N \to \mathbb{R} : (x_i)_{1 \leqslant i \leqslant N} \mapsto \sum_{i=1}^{N} f_i(x_i)$$

and

$$f_i : \zeta \mapsto \begin{cases} |\zeta| - \frac{\mu}{2}, & \text{if } |\zeta| > \mu; \\ \frac{|\zeta|^2}{2\mu}, & \text{if } |\zeta| \leqslant \mu. \end{cases}$$

Note that, since

$$\partial f_i = \nabla f_i : \zeta \mapsto \begin{cases} \frac{\zeta}{|\zeta|}, & \text{if } |\zeta| > \mu; \\ \frac{\zeta}{\mu}, & \text{if } |\zeta| \leqslant \mu, \end{cases}$$

Optimization
Part IV: Proximity operator

Nelly Pustelnik

CNRS, Laboratoire de Physique de l'ENS de Lyon, France

Motivations

Problem: Let $f \in \Gamma_0(\mathbb{R}^N)$, find $\quad \widehat{x} \in \underset{x \in \mathbb{R}^N}{\text{Argmin}}\, f(x)$.

- If f is α-Lipschitz differentiable with $\alpha > 0$, the (explicit) **gradient method**:

$$(\forall n \in \mathbb{N}) \qquad x_{n+1} = x_n - \gamma_n \nabla f(x_n)$$

$\rightarrow$ Convergence insured when $0 < \inf_{n \in \mathbb{N}} \gamma_n$ et $\sup_{n \in \mathbb{N}} \gamma_n < 2\alpha^{-1}$.

- If f nonsmooth, the (explicit) **subgradient method**:

$$(\forall n \in \mathbb{N}) \qquad x_{n+1} = x_n - \gamma_n u_n \quad \text{with} \quad u_n \in \partial f(x_n)$$

$\rightarrow$ Convergence insured when $\gamma_n \in]0, +\infty[$ such that $\sum_{n=0}^{+\infty} \gamma_n^2 < +\infty$ and $\sum_{n=0}^{+\infty} \gamma_n = +\infty$. [Shor, 1979].

- If f nonsmooth, the **implicit subgradient method** is

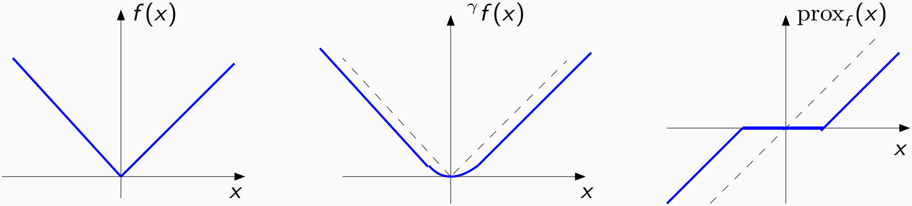
$$(\forall n \in \mathbb{N}) \qquad x_{n+1} = x_n - \gamma_n u_n \quad \text{with} \quad u_n \in \partial f(x_{n+1})$$

$\rightarrow$ Convergence insured when $\sum_{n=0}^{+\infty} \gamma_n = +\infty$ $\Rightarrow$ **Proximity operator**.

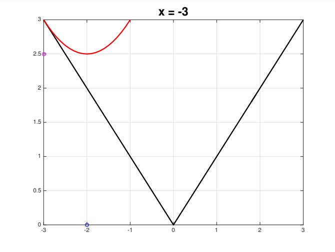
Proximity operator: definition

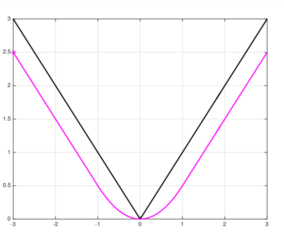
Let $\mathcal{H}$ be a Hilbert space. Let $f \in \Gamma_0(\mathcal{H})$.

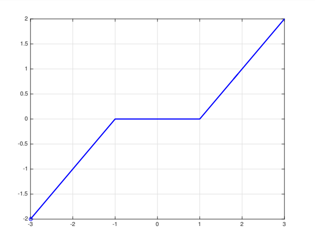
- The **Moreau envelope** of f of parameter $\gamma \in \,]0,+\infty[$ is
$$^{\gamma}f : \mathcal{H} \to \mathbb{R} : x \mapsto \inf_{y\in\mathcal{H}} f(y) + \frac{1}{2\gamma}\|y-x\|^2.$$
- The **proximity operator** of f is
$$\mathrm{prox}_f : \mathcal{H} \to \mathcal{H} : x \mapsto \operatorname*{argmin}_{y\in\mathcal{H}} \; f(y) + \frac{1}{2}\|y-x\|^2.$$



Proximity operator: definition







$f(y) = |y|$

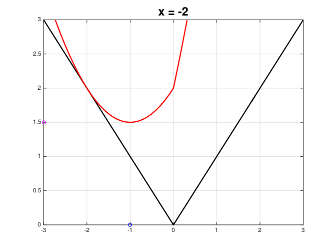
$f(x) = |x|$

$\mathrm{prox}_f(x) = \underset{y \in \mathcal{H}}{\operatorname{argmin}} \; g(y;x)$

$g(y;x) = |y| + \frac{1}{2}(y - x)^2$

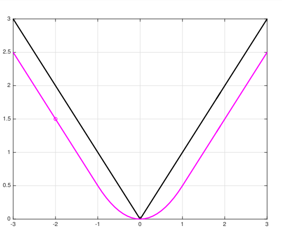
${}^{\gamma}f(x) = \inf_{y \in \mathcal{H}} g(y;x)$

Proximity operator: definition



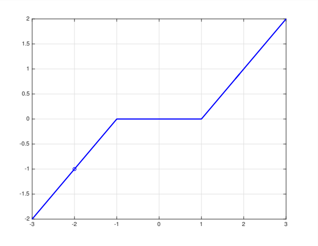
$f(y) = |y|$

$g(y;x) = |y| + \frac{1}{2}(y-x)^2$



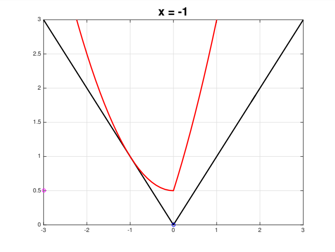
$f(x) = |x|$

${}^{\gamma}f(x) = \inf_{y \in \mathcal{H}} g(y;x)$



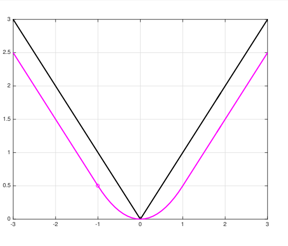
$\mathrm{prox}_f(x) = \operatorname{argmin}_{y \in \mathcal{H}} \; g(y;x)$

Proximity operator: definition



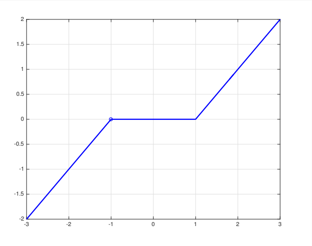
$f(y) = |y|$

$g(y; x) = |y| + \frac{1}{2}(y - x)^2$



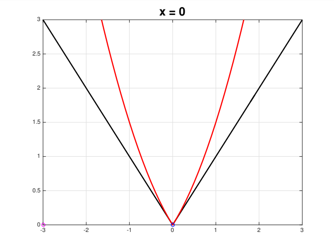
$f(x) = |x|$

${}^{\gamma} f(x) = \inf_{y \in \mathcal{H}} g(y; x)$



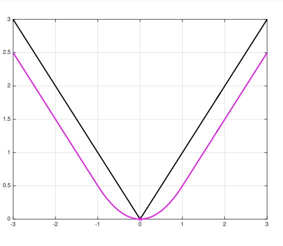
$\operatorname{prox}_f(x) = \operatorname{argmin}_{y \in \mathcal{H}} \; g(y; x)$

Proximity operator: definition



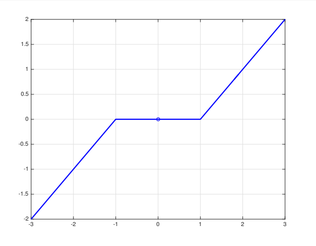
$f(y) = |y|$

$g(y;x) = |y| + \frac{1}{2}(y-x)^2$



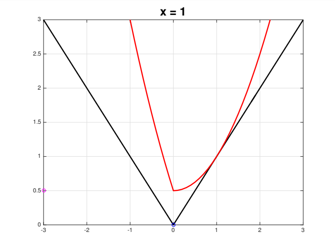
$f(x) = |x|$

${}^{\gamma}f(x) = \inf_{y \in \mathcal{H}} g(y;x)$



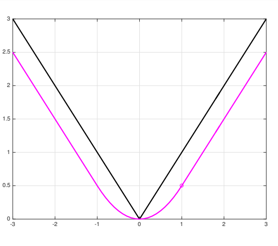
$\mathrm{prox}_f(x) = \operatorname{argmin}_{y \in \mathcal{H}} \; g(y;x)$

Proximity operator: definition



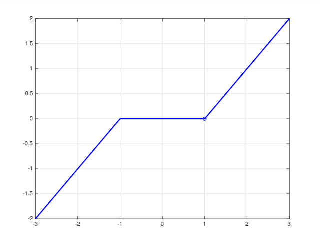
$f(y) = |y|$

$g(y;x) = |y| + \frac{1}{2}(y-x)^2$



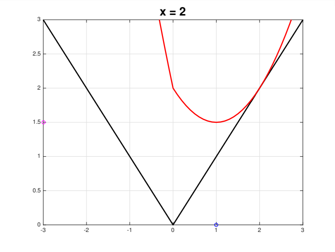
$f(x) = |x|$

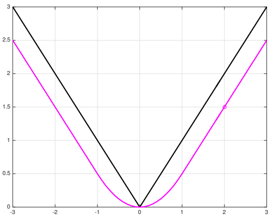
$^{\gamma} f(x) = \inf_{y \in \mathcal{H}} g(y;x)$

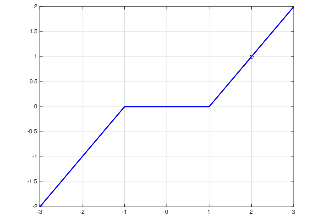


$\mathrm{prox}_f(x) = \operatorname{argmin}_{y \in \mathcal{H}} \; g(y;x)$

Proximity operator: definition







$f(y) = |y|$

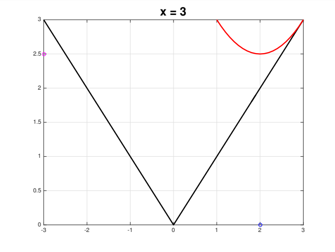
$f(x) = |x|$

$\operatorname{prox}_f(x) = \underset{y \in \mathcal{H}}{\operatorname{argmin}} \; g(y; x)$

$g(y; x) = |y| + \frac{1}{2}(y - x)^2$

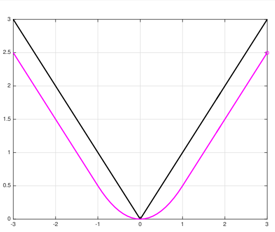
${}^{\gamma} f(x) = \inf_{y \in \mathcal{H}} g(y; x)$

Proximity operator: definition



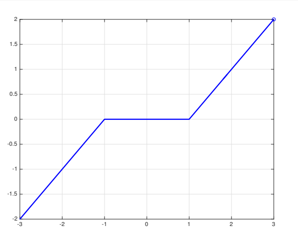
$f(y) = |y|$

$g(y; x) = |y| + \frac{1}{2}(y - x)^2$



$f(x) = |x|$

$^{\gamma}f(x) = \inf_{y \in \mathcal{H}} g(y; x)$



$\mathrm{prox}_f(x) = \underset{y \in \mathcal{H}}{\mathrm{argmin}} \; g(y; x)$

Proximity operator: characterization

Let $\mathcal{H}$ be a Hilbert space and $f \in \Gamma_0(\mathcal{H})$.

$$(\forall x \in \mathcal{H}) \quad p = \mathrm{prox}_f(x) \quad \Leftrightarrow \quad x - p \in \partial f(p).$$

Proximity operator: characterization

Let $\mathcal{H}$ be a Hilbert space and $f \in \Gamma_0(\mathcal{H})$.

$$(\forall x \in \mathcal{H}) \quad p = \mathrm{prox}_f(x) \quad \Leftrightarrow \quad x - p \in \partial f(p).$$

- Proof: By using Fermat's rule, for every $x \in \mathcal{H}$, $p = \mathrm{prox}_f(x)$ if and only if

$$\begin{aligned} & p = \operatorname*{arg\,min}_{y \in \mathcal{H}} \; f(y) + \frac{1}{2}\|y - x\|^2 \\ \Leftrightarrow \quad & 0 \in \partial\Big(f + \frac{1}{2}\|\cdot - x\|^2\Big)(p) \\ \Leftrightarrow \quad & 0 \in \partial f(p) + p - x \end{aligned}$$

Proximity operator: characterization

Let $\mathcal{H}$ be a Hilbert space and $f \in \Gamma_0(\mathcal{H})$.

$$(\forall x \in \mathcal{H}) \quad p = \mathrm{prox}_f(x) \quad \Leftrightarrow \quad x - p \in \partial f(p).$$

- Proof: By using Fermat's rule, for every $x \in \mathcal{H}$, $p = \mathrm{prox}_f(x)$ if and only if

$$\begin{aligned} & p = \underset{y \in \mathcal{H}}{\arg\min}\; f(y) + \frac{1}{2}\|y - x\|^2 \\ \Leftrightarrow \quad & 0 \in \partial\Big(f + \frac{1}{2}\|\cdot - x\|^2\Big)(p) \\ \Leftrightarrow \quad & 0 \in \partial f(p) + p - x \end{aligned}$$

- **Proximal step** :

$$x_{k+1} = \mathrm{prox}_{\gamma f}(x_k) \quad \Leftrightarrow \quad x_{k+1} = x_k - u_k \text{ where } u_k \in \gamma \partial f(x_{k+1})$$

Proximity operator: examples

Projection :

Let $\mathcal{H}$ be a Hilbert space. Let C be a nonempty closed convex subset of $\mathcal{H}$.

$$(\forall x \in \mathcal{H}) \qquad \mathrm{prox}_{\iota_C}(x) = \underset{y \in C}{\mathrm{argmin}} \frac{1}{2}\|y - x\|^2 = P_C(x).$$

Proximity operator: examples

Power q function with $q \geq 1$:

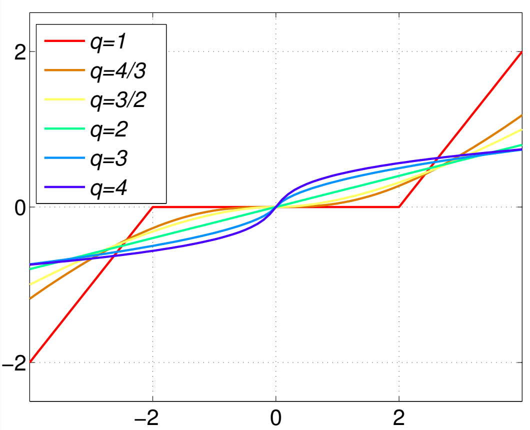
Let $\chi > 0$, $q \in [1, +\infty[$ and $\varphi\colon \mathbb{R} \to\ ]-\infty, +\infty] : \eta \mapsto \chi|\xi|^q$.

Then, for every $\xi \in \mathbb{R}$,

$$\operatorname{prox}_\varphi \xi = \begin{cases} \operatorname{sign}(\xi)\max\{|\xi| - \chi, 0\} & \text{if } q = 1 \\ \xi + \frac{4\chi}{3\,.\,2^{1/3}}\left((\epsilon - \xi)^{1/3} - (\epsilon + \xi)^{1/3}\right) & \\ \qquad \text{where} \quad \epsilon = \sqrt{\xi^2 + 256\chi^3/729} & \text{if } q = \frac{4}{3} \\ \xi + \frac{9\chi^2 \operatorname{sign}(\xi)}{8}\left(1 - \sqrt{1 + \frac{16|\xi|}{9\chi^2}}\right) & \text{if } q = \frac{3}{2} \\ \frac{\xi}{1+2\chi} & \text{if } q = 2 \\ \operatorname{sign}(\xi)\frac{\sqrt{1+12\chi|\xi|}-1}{6\chi} & \text{if } q = 3 \\ \left(\frac{\epsilon+\xi}{8\chi}\right)^{1/3} - \left(\frac{\epsilon-\xi}{8\chi}\right)^{1/3} \quad \text{where} \quad \epsilon = \sqrt{\xi^2 + 1/(27\chi)} & \text{if } q = 4 \end{cases}$$

Proximity operator: examples

Power q function with $q \geq 1$ and $\chi = 2$.



Proximity operator: examples

Quadratic function :

[Combettes-Pesquet, 2010]

Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces.

Let $L \in \mathcal{B}(\mathcal{G}, \mathcal{H})$, $\gamma \in \,]0, +\infty[$ and $z \in \mathcal{G}$.

$$f = \gamma \left\| L \cdot - z \right\|^2 / 2 \quad \Rightarrow \quad \mathrm{prox}_f = (\mathrm{Id} + \gamma L^* L)^{-1}(\cdot + \gamma L^* z).$$

Proximity operator: properties

Let $\mathcal{H}$ be a Hilbert space, $x \in \mathcal{H}$ and $f \in \Gamma_0(\mathcal{H})$.
[Combettes-Pesquet, 2010]

Properties	$g(x)$	$\mathrm{prox}_g x$
Translation	$f(x-z),\ z \in \mathcal{H}$	$z + \mathrm{prox}_f(x-z)$
Quadratic perturbation	$f(x) + \alpha \parallel x \parallel^2 /2 + \langle z \mid x\rangle + \gamma$ $z \in \mathcal{H},\ \alpha > 0,\ \gamma \in \mathbb{R}$	$\mathrm{prox}_{\frac{f}{\alpha+1}}(\frac{x-z}{\alpha+1})$
Scaling	$f(\rho x),\ \rho \in \mathbb{R}^*$	$\frac{1}{\rho}\mathrm{prox}_{\rho^2 f}(\rho x)$
Reflexion	$f(-x)$	$-\mathrm{prox}_f(-x)$
Moreau enveloppe	${}^{\gamma}f(x) = \inf_{y\in\mathcal{H}} f(y) + \frac{1}{2\gamma}\|x-y\|^2$ $\gamma > 0$	$\frac{1}{1+\gamma}\left(\gamma x + \mathrm{prox}_{(1+\gamma)f}(x)\right)$

Proximity operator: properties

For every $i \in \{1, \dots, n\}$, let $\mathcal{H}_i$ be a Hilbert space and let $f_i \in \Gamma_0(\mathcal{H}_i)$.
If

$$\big(\forall x = (x_1, \dots, x_n) \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_n\big) \quad f(x) = \sum_{i=1}^{n} f_i(x_i),$$

then

$$\big(\forall x = (x_1, \dots, x_n) \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_n\big) \quad \mathrm{prox}_f(x) = \big(\mathrm{prox}_{f_i}(x_i)\big)_{1 \le i \le n}.$$

Proximity operator: properties

Let $\mathcal{H}$ be a separable Hilbert space (i.e. if it possesses a countable orthonormal basis).
Let $(b_i)_{i\in I}$ be an orthonormal basis of $\mathcal{H}$.
For every $i \in I$, let $\varphi_i \in \Gamma_0(\mathbb{R})$ such that $\varphi_i \geq 0$. For every $x \in \mathcal{H}$,
if

$$f(x) = \sum_{i\in I} \varphi_i(\langle x \mid b_i \rangle)$$

then

$$\mathrm{prox}_f(x) = \textstyle\sum_{i\in I} \mathrm{prox}_{\varphi_i}(\langle x \mid b_i \rangle) b_i.$$

Remark: The assumption $(\forall i \in I)\ \varphi_i \geq 0$ can be relaxed if $\mathcal{H}$ is finite dimensional.

Proximity operator: properties

Let $\mathcal{H}$ be a separable Hilbert space (i.e. if it possesses a countable orthonormal basis).
Let $(b_i)_{i\in I}$ be an orthonormal basis of $\mathcal{H}$.
For every $i \in I$, let $\varphi_i \in \Gamma_0(\mathbb{R})$ such that $\varphi_i \geq 0$. For every $x \in \mathcal{H}$,
if

$$f(x) = \sum_{i\in I} \varphi_i(\langle x \mid b_i\rangle)$$

then

$$\mathrm{prox}_f(x) = \textstyle\sum_{i\in I} \mathrm{prox}_{\varphi_i}(\langle x \mid b_i\rangle) b_i.$$

Example: $\mathcal{H} = \mathbb{R}^N$, $(b_i)_{1\leq i\leq N}$ canonical basis of $\mathbb{R}^N$, $f = \lambda\|\cdot\|_1$ with $\lambda \in [0,+\infty[$.

$$(\forall x = (x^{(i)})_{1\leq i\leq N} \in \mathbb{R}^N) \qquad \mathrm{prox}_{\lambda\|\cdot\|_1}(x) = \big(\mathrm{prox}_{\lambda|\cdot|}(x^{(i)})\big)_{1\leq i\leq N}$$

Proximity operator: examples

Kullback-Leibler divergence

[Combettes-Pesquet, 2007]

$$(\forall y \in \mathbb{R}^K) \quad f(y;z) = \sum_{k=1}^{K} \phi(y_k)$$

$$\text{where } \phi(y_k) = \begin{cases} -z_k \ln(y_k) + \alpha y_k & \text{if } y_k > 0 \text{ and } z_k > 0 \\ \alpha y_k & \text{if } y_k \geq 0 \text{ and } z_k = 0 \\ +\infty & \text{otherwise} \end{cases}$$

The associated proximity operator is

$$\mathrm{prox}_{\gamma\phi}(y_k) = \frac{y_k - \gamma\alpha + \sqrt{|y_k - \gamma\alpha|^2 + 4\gamma z_k}}{2}$$

Proximity operator: examples

Huber loss

[Combettes-Glaudin, 2019]

$$h : \mathbb{R}^K \to \mathbb{R} : (y_i)_{1\leq i\leq K} \mapsto \sum_{i=1}^{K} h_i(\zeta_i)$$

where

$$h_i : \zeta \mapsto \begin{cases} |\zeta| - \frac{\mu}{2}, & \text{if } |\zeta| > \mu; \\ \frac{|\zeta|^2}{2\mu}, & \text{if } |\zeta| \leqslant \mu. \end{cases}$$

The proximity operator of h can be computed explicitly via

$$\mathrm{prox}_{\tau h} \colon (\zeta_i)_{1\leq i\leq K} \mapsto (\mathrm{prox}_{\tau\phi}\zeta_i)_{1\leq i\leq K}$$

for some $\tau > 0$, where

$$\mathrm{prox}_{\tau\phi} \colon \zeta \mapsto \begin{cases} \zeta - \frac{\tau\zeta}{|\zeta|}, & \text{if } |\zeta| > \tau + \mu; \\ \frac{\mu\zeta}{\tau+\mu}, & \text{if } |\zeta| \leqslant \tau + \mu, \end{cases}$$

Proximity operator: properties

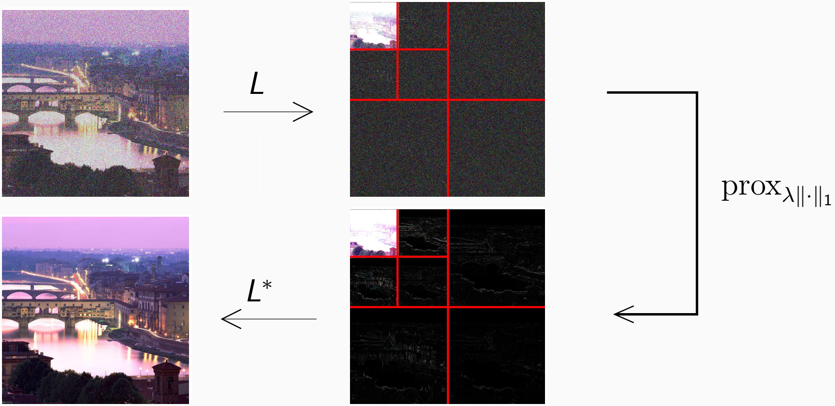
Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces. Let $f \in \Gamma_0(\mathcal{H})$ and $L \in \mathcal{B}(\mathcal{G}, \mathcal{H})$ such that $LL^* = \mu\mathrm{Id}$ where $\mu \in \,]0, +\infty[$. Then

$$\mathrm{prox}_{f \circ L} = \mathrm{Id} - \mu^{-1} L^* \circ (\mathrm{Id} - \mathrm{prox}_{\mu f}) \circ L.$$

Proximity operator: properties

Particular case : $L \in \mathcal{B}(\mathcal{H}, \mathcal{H})$ unitary, $\mathrm{prox}_{f \circ L} = L^* \mathrm{prox}_f L$.

- Illustration: denoising using an ℓ_1 penalty on the coefficients resulting from an orthogonal wavelet transform L.



Proximity operator: closed form expression

- $\mathrm{prox}_{\lambda\|\cdot\|_1}$: soft-thresholding with a fixed threshold $\lambda > 0$.
- $\mathrm{prox}_{\|\cdot\|_{1,2}}$[Peyré,Fadili,2011].
- $\mathrm{prox}_{\|\|_p^p}$ with $p = \{\frac{4}{3}, \frac{3}{2}, 2, 3, 4\}$[Chaux et al.,2005].
- $\mathrm{prox}_{D_{KL}}$[Combettes,Pesquet,2007].
- $\mathrm{prox}_{\iota_C} = P_C$ projection onto the convex set C.

  $\rightarrow$ range constraint hypercube projection,

  $\rightarrow$ $\ell_{1,p}$-ball constraint [Quattoni,2007] [VanDenBerg,2008]
- $\mathrm{prox}_{\sum_{g\in\mathcal{G}} \|\cdot\|_q}$ with overlapping groups [Jenatton et al., 2011]
- Composition with a linear operator: $\mathrm{prox}_{\varphi\circ L}$ closed form if $LL^* = \nu\mathrm{Id}$ [Pustelnik et al., 2016]

Proximity operator: closed form expression

$$\mathrm{prox}_{\varphi_1+\varphi_2} = \mathrm{prox}_{\varphi_2} \circ \mathrm{prox}_{\varphi_1}$$

- [Combettes-Pesquet, 2007] $N = 1$, $\varphi_2 = \iota_C$ of a non-empty closed convex subset of C and φ_1 is differentiable at 0 with $h'(0) = 0$.
- [Chaux-Pesquet-Pustelnik,2009] C and φ_2 are separable in the same basis.
- [Yu, 2013][Shi et al., 2017] $\partial\varphi_2(x) \subset \partial\varphi_2(\mathrm{prox}\varphi_1(x))$.
- Other recent results [Pustelnik, Condat, 2017][Yukawa, Kagami, 2017][del Aguila Pla, Jaldén, 2017]

Useful websites

- Exhaustive list of proximity operators, Matlab and Python codes:
  `http://proximity-operator.net/`
  authors: Chierchia, Chouzenoux, Combettes, Pesquet
- On Github: `https://github.com/cvxgrp/proximal`
  authors: Parikh, Chu, Boyd
- SPAMS: `http://spams-devel.gforge.inria.fr/`
  authors: Mairal, Bach, Ponce, Sapiro, Jenatton, Obozinski
- Fast implementation:
  `https://www.gipsa-lab.grenoble-inp.fr/~laurent.condat/software.html`
  author: Condat

Optimization
Part V: Algorithms

Nelly Pustelnik

CNRS, Laboratoire de Physique de l'ENS de Lyon, France

(part written in collaboration with **Luis Briceño-Arias** from Department of Mathematics, Universidad Técnica Federico Santa Marìa, Santiago, Chile)

Fixed point algorithm

Fixed point algorithm: zeros and fixed points

$2^{\mathcal{H}}$ is the power set of $\mathcal{H}$, i.e. the family of all subsets of $\mathcal{H}$.

Let $\mathcal{H}$ be a Hilbert space. Let $\Phi\colon \mathcal{H} \to 2^{\mathcal{H}}$.

The set of **fixed points** of Φ is : $\mathrm{Fix}\Phi = \{x \in \mathcal{H} \,|\, x \in \Phi x\}$.

The set of **zeros** of Φ is : $\mathrm{zer}\Phi = \{x \in \mathcal{H} \,|\, 0 \in \Phi x\}$.

Fixed point algorithm: convergence

Let $\mathcal{H}$ be a Hilbert space.

Let $(x_k)_{k\in\mathbb{N}}$ be a sequence in $\mathcal{H}$ and $\widehat{x} \in \mathcal{H}$.

- $(x_k)_{k\in\mathbb{N}}$ **converges strongly** to $\widehat{x}$ if

$$\lim_{k\to+\infty} \|x_k - \widehat{x}\| = 0.$$

  It is denoted by $x_k \to \widehat{x}$.

- $(x_k)_{k\in\mathbb{N}}$ **converges weakly** to $\widehat{x}$ if

$$(\forall y \in \mathcal{H}) \qquad \lim_{n\to+\infty} \langle y \mid x_k - \widehat{x}\rangle = 0.$$

  It is denoted by $x_k \rightharpoonup \widehat{x}$.

Remark: In a finite dimensional Hilbert space, strong and weak convergences are equivalent.

Banach-Picard theorem

An operator $\Phi\colon \mathcal{H} \to \mathcal{H}$ is **$\omega-$Lipschitz continuous** for some $\omega \in [0, +\infty[$ if

$$(\forall x \in \mathcal{H})(\forall y \in \mathcal{H}) \quad \|\Phi x - \Phi y\| \leq \omega \|x - y\|.$$

Φ is **nonexpansive** if it is $1-$Lipschitz continuous.

Banach-Picard theorem Let $\omega \in [0, 1[$, let $\Phi\colon \mathcal{H} \to \mathcal{H}$ be a $\omega-$Lipschitz continuous operator, and let $x_0 \in \mathcal{H}$. Set

$$(\forall k \in \mathbb{N}) \quad x_{k+1} = \Phi x_k.$$

Then, $\mathrm{Fix}\Phi = \{\widehat{x}\}$ for some $\widehat{x} \in \mathcal{H}$ and we have

$$(\forall k \in \mathbb{N}) \quad \|x_k - \widehat{x}\| \leq \omega^k \|x_0 - \widehat{x}\|.$$

Moreover, $(x_k)_{k\in\mathbb{N}}$ converges strongly to $\widehat{x}$ with linear convergence rate ω.

Gâteaux differentiable + Lipschitz

An operator $\Phi\colon \mathcal{H} \to \mathcal{H}$ is $\omega-$**Lipschitz continuous** for some $\omega \in [0,+\infty[$ if

$$(\forall x \in \mathcal{H})(\forall y \in \mathcal{H}) \quad \|\Phi x - \Phi y\| \leq \omega \|x - y\|.$$

For every $\zeta \geq 0$, we consider the class $C^{1,1}_{\zeta}(\mathcal{H})$ of functions $f\colon \mathcal{H} \to \mathbb{R}$ satisfying:

- f is Gâteaux differentiable in $\mathcal{H}$
- $\nabla f\colon \mathcal{H} \to \mathcal{H}$ is ζ-Lipschitz continuous.

Averaged nonexpansive operator

An operator $\Phi\colon \mathcal{H} \to \mathcal{H}$ is $\mu-$**averaged nonexpansive** for some $\mu \in]0,1]$ if, for every $x \in \mathcal{H}$ and $y \in \mathcal{H}$,

$$\|\Phi x - \Phi y\|^2 \leq \|x - y\|^2 - \left(\frac{1-\mu}{\mu}\right) \|(\mathrm{Id} - \Phi)x - (\mathrm{Id} - \Phi)y\|^2,$$

Φ is **firmly nonexpansive** if it is $1/2-$averaged.

Φ is **nonexpansive** if and only if Φ is $1-$averaged.

Theorem Let $\mu \in]0,1[$, let $\Phi\colon \mathcal{H} \to \mathcal{H}$ be a $\mu-$averaged nonexpansive operator such that $\mathrm{Fix}\Phi \neq \varnothing$, and let $x_0 \in \mathcal{H}$. Set

$$(\forall k \in \mathbb{N}) \quad x_{k+1} = \Phi x_k.$$

Then $(x_k)_{k\in\mathbb{N}}$ converges weakly to a point in $\mathrm{Fix}\Phi$.

Cocoercive and strongly monotone operators

For every $\eta \in [0, +\infty[$, we define the class $\mathcal{C}_\eta$ of $\eta-$**cocoercive operators** $\mathcal{M}\colon \mathcal{H} \to \mathcal{H}$ satisfying, for every x and y in $\mathcal{H}$,

$$\langle \mathcal{M}x - \mathcal{M}y \mid x - y \rangle \geq \eta \|\mathcal{M}x - \mathcal{M}y\|^2.$$

If $\mathcal{M} \in \mathcal{C}_\eta$ for some $\eta > 0$, then $\mathcal{M}$ is $\eta^{-1}-$Lipschitz continuous

Let $\zeta \geq 0$ and let $f\colon \mathcal{H} \to \mathbb{R}$ be a convex function. Then the following are equivalent:

1. $f \in C^{1,1}_\zeta(\mathcal{H})$ (i.e. Gâteaux differentiable + ζ-Lipschitz continuous).
2. f is Fréchet differentiable and $\nabla f \in \mathcal{C}_{1/\zeta}$.

Strongly monotone operators

An operator $\mathcal{M}\colon \mathcal{H} \to \mathcal{H}$ is $\rho-$**strongly monotone** for some $\rho \in \ ]0, +\infty[$ if, for every x and y in $\mathcal{H}$,

$$\langle \mathcal{M}x - \mathcal{M}y \mid x - y \rangle \geq \rho \|x - y\|^2.$$

A function $h \in C_{\zeta}^{1,1}(\mathcal{H})$ is $\rho-$strongly convex, for some $\rho \in \ ]0, +\infty[$, if $h - \frac{\rho}{2}\|\cdot\|_2^2$ is convex or, equivalently, if ∇h is $\rho-$strongly monotone.

Nonlinear operators

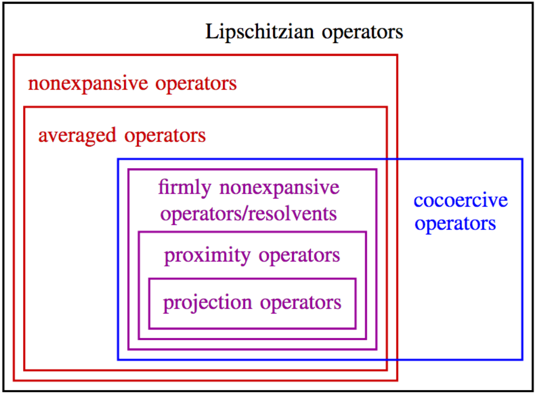


Fig. 3: Classes of nonlinear operators.

→ extracted from [Combettes-Pesquet, 2020]

Gradient method

Let $f \in \Gamma_0(\mathcal{H})$ and $f \in C^{1,1}_{\zeta}(\mathcal{H})$ (i.e. Gâteaux differentiable + ζ-Lipschitz continuous). We set, for some $\tau > 0$,

$$\Phi := \mathrm{Id} - \tau \nabla f$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = x_k - \tau \nabla f(x_k)$.

Gradient method

Let $f \in \Gamma_0(\mathcal{H})$ and $f \in C_\zeta^{1,1}(\mathcal{H})$ (i.e. Gâteaux differentiable + ζ-Lipschitz continuous). We set, for some $\tau > 0$,

$$\Phi := \mathrm{Id} - \tau \nabla f$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = x_k - \tau \nabla f(x_k)$.
- For every $\tau > 0$, $\mathrm{zer}\, \nabla f = (\nabla f)^{-1}(0) = \mathrm{Fix}\Phi$.

Gradient method

Let $f \in \Gamma_0(\mathcal{H})$ and $f \in C^{1,1}_{\zeta}(\mathcal{H})$ (i.e. Gâteaux differentiable + ζ-Lipschitz continuous). We set, for some $\tau > 0$,

$$\Phi := \mathrm{Id} - \tau \nabla f$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = x_k - \tau \nabla f(x_k)$.
- For every $\tau > 0$, $\operatorname{zer} \nabla f = (\nabla f)^{-1}(0) = \mathrm{Fix}\Phi$.
- For every $\nabla f \in \mathcal{C}_{1/\zeta}$ (i.e. $1/\zeta-$cocoercive operator) and every $\tau \in \left]0, 2\zeta^{-1}\right[$, Φ is averaged nonexpansive.

  $\rightarrow$ cf. Proposition 4.39 in [Bauschke-Combettes, 2017]

Gradient method

Let $f \in \Gamma_0(\mathcal{H})$ and $f \in C_\zeta^{1,1}(\mathcal{H})$ (i.e. Gâteaux differentiable + ζ-Lipschitz continuous). We set, for some $\tau > 0$,

$$\Phi := \mathrm{Id} - \tau \nabla f$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = x_k - \tau \nabla f(x_k)$.
- For every $\tau > 0$, $\mathrm{zer}\, \nabla f = (\nabla f)^{-1}(0) = \mathrm{Fix}\Phi$.
- For every $\nabla f \in \mathcal{C}_{1/\zeta}$ (i.e. $1/\zeta-$cocoercive operator) and every $\tau \in \left]0, 2\zeta^{-1}\right[$, Φ is averaged nonexpansive.

  $\rightarrow$ cf. Proposition 4.39 in [Bauschke-Combettes, 2017]
- The gradient method **converges weakly** to a point in $\mathrm{zer}\, \nabla f$.

  $\rightarrow$ cf. p.6

Gradient method

Let $f \in \Gamma_0(\mathcal{H})$ and $f \in C^{1,1}_\zeta(\mathcal{H})$ (i.e. Gâteaux differentiable + ζ-Lipschitz continuous). We set, for some $\tau > 0$,

$$\Phi := \mathrm{Id} - \tau \nabla f$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = x_k - \tau \nabla f(x_k)$.
- For every $\tau > 0$, $\operatorname{zer} \nabla f = (\nabla f)^{-1}(0) = \mathrm{Fix}\Phi$.
- For every $\nabla f \in \mathcal{C}_{1/\zeta}$ (i.e. $1/\zeta-$cocoercive operator) and every $\tau \in \left]0, 2\zeta^{-1}\right[$, Φ is averaged nonexpansive.

  $\rightarrow$ cf. Proposition 4.39 in [Bauschke-Combettes, 2017]
- The gradient method **converges weakly** to a point in $\operatorname{zer} \nabla f$.

  $\rightarrow$ cf. p.6
- If ∇f is additionally strongly monotone and $\tau \in \left]0, 2\zeta^{-1}\right[$, Φ is Lipschitz continuous with constant in $]0,1[$ and EA achieves **linear convergence**.

  $\rightarrow$ Fact 7 in [Yin et al., 2020]

Proximal Point Algorithm (PPA)

Let $f \in \Gamma_0(\mathcal{H})$. We set, for some $\tau > 0$,

$$\Phi := \mathrm{prox}_{\tau f} = (\mathrm{Id} + \tau\partial f)^{-1}.$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = \mathrm{prox}_{\gamma f}(x_k)$.

Proximal Point Algorithm (PPA)

Let $f \in \Gamma_0(\mathcal{H})$. We set, for some $\tau > 0$,

$$\Phi := \mathrm{prox}_{\tau f} = (\mathrm{Id} + \tau \partial f)^{-1}.$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = \mathrm{prox}_{\gamma f}(x_k)$.
- For every $\tau > 0$, $\mathrm{Fix}\Phi = \mathrm{zer}\, \partial f$.

  Proof:

$$\begin{aligned} x = \mathrm{prox}_{\tau f} x &\Leftrightarrow x \in (\mathrm{I} + \tau \partial f) x \\ &\Leftrightarrow x \in x + \tau \partial f(x) \\ &\Leftrightarrow 0 \in \partial f \end{aligned}$$

Proximal Point Algorithm (PPA)

Let $f \in \Gamma_0(\mathcal{H})$. We set, for some $\tau > 0$,

$$\Phi := \mathrm{prox}_{\tau f} = (\mathrm{Id} + \tau\partial f)^{-1}.$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = \mathrm{prox}_{\gamma f}(x_k)$.
- For every $\tau > 0$, $\mathrm{Fix}\Phi = \mathrm{zer}\,\partial f$.
- For every $\tau > 0$ and any $f \in \Gamma_0(\mathcal{H})$, $\mathrm{prox}_{\tau f}$ is firmly nonexpansive.

  $\rightarrow$ cf. Proposition 23.8 in [Bauschke-Combettes, 2017]

Proximal Point Algorithm (PPA)

Let $f \in \Gamma_0(\mathcal{H})$. We set, for some $\tau > 0$,

$$\Phi := \operatorname{prox}_{\tau f} = (\mathrm{Id} + \tau \partial f)^{-1}.$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = \operatorname{prox}_{\gamma f}(x_k)$.
- For every $\tau > 0$, $\mathrm{Fix}\Phi = \mathrm{zer}\, \partial f$.
- For every $\tau > 0$ and any $f \in \Gamma_0(\mathcal{H})$, $\operatorname{prox}_{\tau f}$ is firmly nonexpansive.

  $\rightarrow$ cf. Proposition 23.8 in [Bauschke-Combettes, 2017]
- The **PPA method converges weakly** to a point in $\mathrm{zer}\, \partial f$.

  $\rightarrow$ cf. p.6

Proximal Point Algorithm (PPA)

Let $f \in \Gamma_0(\mathcal{H})$. We set, for some $\tau > 0$,

$$\Phi := \mathrm{prox}_{\tau f} = (\mathrm{Id} + \tau \partial f)^{-1}.$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = \mathrm{prox}_{\gamma f}(x_k)$.
- For every $\tau > 0$, $\mathrm{Fix}\Phi = \mathrm{zer}\,\partial f$.
- For every $\tau > 0$ and any $f \in \Gamma_0(\mathcal{H})$, $\mathrm{prox}_{\tau f}$ is firmly nonexpansive.

  $\rightarrow$ cf. Proposition 23.8 in [Bauschke-Combettes, 2017]
- The **PPA method converges weakly** to a point in $\mathrm{zer}\,\partial f$.

  $\rightarrow$ cf. p.6
- If ∂f is additionally strongly monotone we obtain that $\mathrm{prox}_{\tau f}$ is Lipschitz continuous with constant in $]0, 1[$ and **PPA achieves linear convergence.**

  $\rightarrow$ Proposition 23.13 in [Bauschke-Combettes, 2017]

α-averaged operator: example

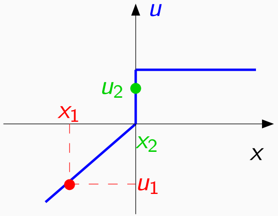
Let $\mathcal{H}$ be a Hilbert space, $f \in \Gamma_0(\mathcal{H})$.
prox_f is a $1/2$-averaged operator.

α-averaged operator: example

Let $\mathcal{H}$ be a Hilbert space, $f \in \Gamma_0(\mathcal{H})$.
prox_f is a $1/2$-averaged operator.

Proof:

- We recall that : $\partial f(x) = \{u \in \mathcal{H} \,|\, (\forall y \in \mathcal{H})\ \langle y - x | u \rangle + f(x) \leq f(y)\}$



α-averaged operator: example

Let $\mathcal{H}$ be a Hilbert space, $f \in \Gamma_0(\mathcal{H})$.
prox_f is a 1/2-averaged operator.

Proof:

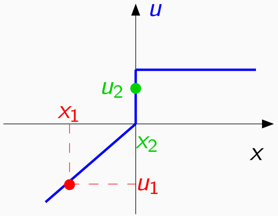
- We recall that : $\partial f(x) = \{u \in \mathcal{H} \,|\, (\forall y \in \mathcal{H})\ \langle y - x | u \rangle + f(x) \leq f(y)\}$

- Let $u_1 \in \partial f(x_1)$ and $u_2 \in \partial f(x_2)$.

  By definition:

$$\langle x_2 - x_1 | u_1 \rangle + f(x_1) \leq f(x_2)$$
$$\langle x_1 - x_2 | u_2 \rangle + f(x_2) \leq f(x_1)$$

  it results that $\boxed{\langle x_1 - x_2 \,|\, u_1 - u_2 \rangle \geq 0}$.



α-averaged operator: example

Let $\mathcal{H}$ be a Hilbert space, $f \in \Gamma_0(\mathcal{H})$.
prox_f is a $1/2$-averaged operator.

Proof:

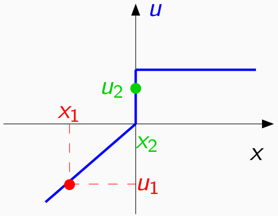
- We recall that : $\partial f(x) = \{u \in \mathcal{H} \,|\, (\forall y \in \mathcal{H})\ \langle y - x | u \rangle + f(x) \leq f(y)\}$

- Let $u_1 \in \partial f(x_1)$ and $u_2 \in \partial f(x_2)$.

  By definition:

$$\langle x_2 - x_1 | u_1 \rangle + f(x_1) \leq f(x_2)$$
$$\langle x_1 - x_2 | u_2 \rangle + f(x_2) \leq f(x_1)$$

  it results that $\boxed{\langle x_1 - x_2 \,|\, u_1 - u_2 \rangle \geq 0}$.



α-averaged operator: example

Proof :

Let $u_1 \in \partial f(x_1)$ and $u_2 \in \partial f(x_2)$

$$\langle x_1 - x_2 \mid u_1 - u_2 \rangle \geq 0 \Leftrightarrow \quad \langle x_1 - x_2 \mid x_1 - x_2 + u_1 - u_2 \rangle \geq \| x_1 - x_2 \|^2$$

We consider $u_1' \in (\mathrm{Id} + \partial f)x_1$ et $u_2' \in (\mathrm{Id} + \partial f)x_2$, it results that

$$\langle x_1 - x_2 \mid u_1' - u_2' \rangle \geq \| x_1 - x_2 \|^2$$

Then, from the definition of the proximity operator,

$$\langle \mathrm{prox}_f u_1' - \mathrm{prox}_f u_2' \mid u_1' - u_2' \rangle \geq \| \mathrm{prox}_f u_1' - \mathrm{prox}_f u_2' \|^2$$

We can deduce that prox_f is a 1/2-averaged operator, i.e,

$$\| u_1' - u_2' \|^2 \geq \| \mathrm{prox}_f u_1' - \mathrm{prox}_f u_2' \|^2 + \| (\mathrm{Id} - \mathrm{prox}_f) u_1' - (\mathrm{Id} - \mathrm{prox}_f) u_2' \|^2$$

Forward-backward splitting

Let $f \in \Gamma_0(\mathcal{H})$ and $g \in \Gamma_0(\mathcal{H})$. Additionally, $f \in C^{1,1}_\zeta(\mathcal{H})$ (i.e. Gâteaux differentiable + ζ-Lipschitz continuous). We set, for some $\tau > 0$,

$$\Phi := \mathrm{prox}_{\tau g}(\mathrm{Id} - \tau \nabla f) = (\mathrm{Id} + \tau \partial g)^{-1}(\mathrm{Id} - \tau \nabla f)$$

- **Iterations**:$(\forall k \in \mathbb{N})$ $\quad x_{k+1} = \mathrm{prox}_{\tau g}(x_k - \tau \nabla f(x_k))$.

Forward-backward splitting

Let $f \in \Gamma_0(\mathcal{H})$ and $g \in \Gamma_0(\mathcal{H})$. Additionally, $f \in C_\zeta^{1,1}(\mathcal{H})$ (i.e. Gâteaux differentiable + ζ-Lipschitz continuous). We set, for some $\tau > 0$,

$$\Phi := \mathrm{prox}_{\tau g}(\mathrm{Id} - \tau \nabla f) = (\mathrm{Id} + \tau \partial g)^{-1}(\mathrm{Id} - \tau \nabla f)$$

- **Iterations**:$(\forall k \in \mathbb{N})$ $\quad x_{k+1} = \mathrm{prox}_{\tau g}(x_k - \tau \nabla f(x_k))$.
- Roots in projected gradient method [Levitin 1966] when $g = \iota_C$ for some closed convex set C

Forward-backward splitting

Let $f \in \Gamma_0(\mathcal{H})$ and $g \in \Gamma_0(\mathcal{H})$. Additionally, $f \in C^{1,1}_\zeta(\mathcal{H})$ (i.e. Gâteaux differentiable + ζ-Lipschitz continuous). We set, for some $\tau > 0$,

$$\Phi := \mathrm{prox}_{\tau g}(\mathrm{Id} - \tau\nabla f) = (\mathrm{Id} + \tau\partial g)^{-1}(\mathrm{Id} - \tau\nabla f)$$

- **Iterations**:$(\forall k \in \mathbb{N}) \quad x_{k+1} = \mathrm{prox}_{\tau g}(x_k - \tau\nabla f(x_k))$.
- For every $\tau > 0$, $\mathrm{zer}\,(\nabla f + \partial g) = \mathrm{Fix}\Phi$.
  Proof:

$$\begin{aligned} x \in \mathrm{Fix}\Phi &\Leftrightarrow (\mathrm{Id} - \gamma\nabla f)x \in (\mathrm{Id} + \gamma\partial g)x \\ &\Leftrightarrow 0 \in \nabla f(x) + \partial g(x). \end{aligned}$$

Forward-backward splitting

Let $f \in \Gamma_0(\mathcal{H})$ and $g \in \Gamma_0(\mathcal{H})$. Additionally, $f \in C^{1,1}_\zeta(\mathcal{H})$ (i.e. Gâteaux differentiable + ζ-Lipschitz continuous). We set, for some $\tau > 0$,

$$\Phi := \mathrm{prox}_{\tau g}(\mathrm{Id} - \tau\nabla f) = (\mathrm{Id} + \tau\partial g)^{-1}(\mathrm{Id} - \tau\nabla f)$$

- **Iterations**:$(\forall k \in \mathbb{N})$ $\quad x_{k+1} = \mathrm{prox}_{\tau g}(x_k - \tau\nabla f(x_k))$.
- For every $\tau > 0$, $\mathrm{zer}\,(\nabla f + \partial g) = \mathrm{Fix}\Phi$.
- $f \in \Gamma_0(\mathcal{H})$, ∇f is $\zeta^{-1}-$cocoercive and $\tau \in \left]0, 2\zeta^{-1}\right[$, $\mathrm{prox}_{\tau g}(\mathrm{Id} - \tau\nabla f)$ is averaged nonexpansive.

  $\rightarrow$ cf. Theorem 26.14, Proposition 26.1(iv)(d) [Bauschke-Combettes, 2017]

Forward-backward splitting

Let $f \in \Gamma_0(\mathcal{H})$ and $g \in \Gamma_0(\mathcal{H})$. Additionally, $f \in C^{1,1}_\zeta(\mathcal{H})$ (i.e. Gâteaux differentiable + ζ-Lipschitz continuous). We set, for some $\tau > 0$,

$$\Phi := \mathrm{prox}_{\tau g}(\mathrm{Id} - \tau\nabla f) = (\mathrm{Id} + \tau\partial g)^{-1}(\mathrm{Id} - \tau\nabla f)$$

- **Iterations**:$(\forall k \in \mathbb{N})$ $\quad x_{k+1} = \mathrm{prox}_{\tau g}(x_k - \tau\nabla f(x_k))$.
- For every $\tau > 0$, $\mathrm{zer}\,(\nabla f + \partial g) = \mathrm{Fix}\Phi$.
- $f \in \Gamma_0(\mathcal{H})$, ∇f is ζ^{-1}−cocoercive and $\tau \in \left]0, 2\zeta^{-1}\right[$, $\mathrm{prox}_{\tau g}(\mathrm{Id} - \tau\nabla f)$ is averaged nonexpansive.

  $\rightarrow$ cf. Theorem 26.14, Proposition 26.1(iv)(d) [Bauschke-Combettes, 2017]
- The **FBS method converges weakly** to a point in $\mathrm{zer}\,(\nabla f + \partial g)$.

  $\rightarrow$ consequence of the averaged nonexpansiveness

Forward-backward splitting

Let $f \in \Gamma_0(\mathcal{H})$ and $g \in \Gamma_0(\mathcal{H})$. Additionally, $f \in C^{1,1}_{\zeta}(\mathcal{H})$ (i.e. Gâteaux differentiable + ζ-Lipschitz continuous). We set, for some $\tau > 0$,

$$\Phi := \mathrm{prox}_{\tau g}(\mathrm{Id} - \tau \nabla f) = (\mathrm{Id} + \tau \partial g)^{-1}(\mathrm{Id} - \tau \nabla f)$$

- **Iterations**:$(\forall k \in \mathbb{N}) \quad x_{k+1} = \mathrm{prox}_{\tau g}(x_k - \tau \nabla f(x_k))$.
- For every $\tau > 0$, $\mathrm{zer}\,(\nabla f + \partial g) = \mathrm{Fix}\Phi$.
- $f \in \Gamma_0(\mathcal{H})$, ∇f is $\zeta^{-1}-$cocoercive and $\tau \in \left]0, 2\zeta^{-1}\right[$, $\mathrm{prox}_{\tau g}(\mathrm{Id} - \tau \nabla f)$ is averaged nonexpansive.

  $\rightarrow$ cf. Theorem 26.14, Proposition 26.1(iv)(d) [Bauschke-Combettes, 2017]
- The **FBS method converges weakly** to a point in $\mathrm{zer}\,(\nabla f + \partial g)$.

  $\rightarrow$ consequence of the averaged nonexpansiveness
- If ∂g or ∇f is additionally strongly monotone we obtain that Φ is Lipschitz continuous with constant in $]0, 1[$ and **FBS achieves linear convergence.**

  $\rightarrow$ Theorem 26.16 in [Bauschke-Combettes, 2017]

Peaceman-Rachford splitting

Let $f \in \Gamma_0(\mathcal{H})$ and $g \in \Gamma_0(\mathcal{H})$. We set, for some $\tau > 0$,

$$\Phi := (2\mathrm{prox}_{\tau g} - \mathrm{Id}) \circ (2\mathrm{prox}_{\tau f} - \mathrm{Id})$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = 2\mathrm{prox}_{\tau g}(2\mathrm{prox}_{\tau f} x_k - x_k) - 2\mathrm{prox}_{\tau f} x_k + x_k.$

Peaceman-Rachford splitting

Let $f \in \Gamma_0(\mathcal{H})$ and $g \in \Gamma_0(\mathcal{H})$. We set, for some $\tau > 0$,

$$\Phi := (2\mathrm{prox}_{\tau g} - \mathrm{Id}) \circ (2\mathrm{prox}_{\tau f} - \mathrm{Id})$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = 2\mathrm{prox}_{\tau g}(2\mathrm{prox}_{\tau f} x_k - x_k) - 2\mathrm{prox}_{\tau f} x_k + x_k$.
- For every $\tau > 0$, $\mathrm{zer}\,(\nabla f + \partial g) = \mathrm{prox}_{\tau f}(\mathrm{Fix}\Phi)$.

  $\rightarrow$ cf. Proposition 26.1(iii)(b)[Bauschke-Combettes, 2017]

Peaceman-Rachford splitting

Let $f \in \Gamma_0(\mathcal{H})$ and $g \in \Gamma_0(\mathcal{H})$. We set, for some $\tau > 0$,

$$\Phi := (2\mathrm{prox}_{\tau g} - \mathrm{Id}) \circ (2\mathrm{prox}_{\tau f} - \mathrm{Id})$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = 2\mathrm{prox}_{\tau g}(2\mathrm{prox}_{\tau f} x_k - x_k) - 2\mathrm{prox}_{\tau f} x_k + x_k$.
- For every $\tau > 0$, $\mathrm{zer}\,(\nabla f + \partial g) = \mathrm{prox}_{\tau f}(\mathrm{Fix}\Phi)$.

  $\rightarrow$ cf. Proposition 26.1(iii)(b)[Bauschke-Combettes, 2017]
- Weak **convergence not guaranteed in this general setting**.
- Weak **convergence guaranteed if $f \in C^{1,1}_{\zeta}(\mathcal{H})$.**

  $\rightarrow$ cf. Corollary 1& Remark 2(2) [Lions-Mercier, 1979]

Peaceman-Rachford splitting

Let $f \in \Gamma_0(\mathcal{H})$ and $g \in \Gamma_0(\mathcal{H})$. We set, for some $\tau > 0$,

$$\Phi := (2\mathrm{prox}_{\tau g} - \mathrm{Id}) \circ (2\mathrm{prox}_{\tau f} - \mathrm{Id})$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = 2\mathrm{prox}_{\tau g}(2\mathrm{prox}_{\tau f} x_k - x_k) - 2\mathrm{prox}_{\tau f} x_k + x_k$.
- For every $\tau > 0$, $\mathrm{zer}\,(\nabla f + \partial g) = \mathrm{prox}_{\tau f}(\mathrm{Fix}\Phi)$.

  → cf. Proposition 26.1(iii)(b)[Bauschke-Combettes, 2017]
- Weak **convergence not guaranteed in this general setting**.
- Weak **convergence guaranteed if $f \in C^{1,1}_{\zeta}(\mathcal{H})$.**

  → cf. Corollary 1& Remark 2(2) [Lions-Mercier, 1979]
- If $f \in C^{1,1}_{\zeta}(\mathcal{H})$ and ∇f is additionally strongly monotone we obtain that $2\mathrm{prox}_{\tau f} - \mathrm{Id}$ is Lipschitz continuous with constant in $]0,1[$ and **Φ achieves linear convergence.**

  → [**Giselson, 2017**]

Douglas-Rachford splitting

Let $f \in \Gamma_0(\mathcal{H})$ and $g \in \Gamma_0(\mathcal{H})$. We set, for some $\tau > 0$,

$$\Phi := \mathrm{prox}_{\tau g} \circ (2\mathrm{prox}_{\tau f} - \mathrm{Id}) + \mathrm{Id} - \mathrm{prox}_{\tau f}$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = \mathrm{prox}_{\tau g}(2\mathrm{prox}_{\tau f} x_k - x_k) + x_k - \mathrm{prox}_{\tau f} x_k$.

Douglas-Rachford splitting

Let $f \in \Gamma_0(\mathcal{H})$ and $g \in \Gamma_0(\mathcal{H})$. We set, for some $\tau > 0$,

$$\Phi := \mathrm{prox}_{\tau g} \circ (2\mathrm{prox}_{\tau f} - \mathrm{Id}) + \mathrm{Id} - \mathrm{prox}_{\tau f}$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = \mathrm{prox}_{\tau g}(2\mathrm{prox}_{\tau f} x_k - x_k) + x_k - \mathrm{prox}_{\tau f} x_k$.
- For every $\tau > 0$, $\mathrm{zer}\,(\nabla f + \partial g) = \mathrm{prox}_{\tau f}(\mathrm{Fix}\Phi)$.

  $\rightarrow$ cf. Proposition 26.1(iii)(b)[Bauschke-Combettes, 2017]

Douglas-Rachford splitting

Let $f \in \Gamma_0(\mathcal{H})$ and $g \in \Gamma_0(\mathcal{H})$. We set, for some $\tau > 0$,

$$\Phi := \mathrm{prox}_{\tau g} \circ (2\mathrm{prox}_{\tau f} - \mathrm{Id}) + \mathrm{Id} - \mathrm{prox}_{\tau f}$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = \mathrm{prox}_{\tau g}(2\mathrm{prox}_{\tau f} x_k - x_k) + x_k - \mathrm{prox}_{\tau f} x_k$.
- For every $\tau > 0$, $\mathrm{zer}\,(\nabla f + \partial g) = \mathrm{prox}_{\tau f}(\mathrm{Fix}\Phi)$.

  $\rightarrow$ cf. Proposition 26.1(iii)(b)[Bauschke-Combettes, 2017]
- **Weak convergence** insured as

  $\rightarrow$ cf. Corollary 1& Remark 2(2) [Lions-Mercier, 1979]

Douglas-Rachford splitting

Let $f \in \Gamma_0(\mathcal{H})$ and $g \in \Gamma_0(\mathcal{H})$. We set, for some $\tau > 0$,

$$\Phi := \mathrm{prox}_{\tau g} \circ (2\mathrm{prox}_{\tau f} - \mathrm{Id}) + \mathrm{Id} - \mathrm{prox}_{\tau f}$$

- **Iterations**: $(\forall k \in \mathbb{N}) \quad x_{k+1} = \mathrm{prox}_{\tau g}(2\mathrm{prox}_{\tau f} x_k - x_k) + x_k - \mathrm{prox}_{\tau f} x_k$.
- For every $\tau > 0$, $\mathrm{zer}\,(\nabla f + \partial g) = \mathrm{prox}_{\tau f}(\mathrm{Fix}\Phi)$.

  $\rightarrow$ cf. Proposition 26.1(iii)(b)[Bauschke-Combettes, 2017]
- **Weak convergence** insured as

  $\rightarrow$ cf. Corollary 1& Remark 2(2) [Lions-Mercier, 1979]
- If $f \in C^{1,1}_{\zeta}(\mathcal{H})$ and ∇f is additionally strongly monotone we obtain that $2\mathrm{prox}_{\tau f} - \mathrm{Id}$ is Lipschitz continuous with constant in $]0,1[$ and **Φ achieves linear convergence.**

  $\rightarrow$ [**Giselson, 2017**]

Comparisons

Let $f \in C^{1,1}_{1/\alpha}(\mathcal{H})$ and $g \in C^{1,1}_{1/\beta}(\mathcal{H})$, for some $\alpha > 0$ and $\beta > 0$.
The problem is to

$$\underset{x \in \mathcal{H}}{\text{minimize}} \;\; f(x) + g(x),$$

under the assumption that solutions exist.

Example: **Smooth TV denoising**

$$\underset{x \in \mathbb{R}^N}{\text{minimize}} \;\; \frac{1}{2}\|x - z\|_2^2 + \chi h_\mu(Lx),$$

- $L \in \mathbb{R}^{N-1 \times N}$ denotes the first order discrete difference operator

$$(\forall n \in \{1, \ldots, N-1\}) \quad (Lx)_n = \tfrac{1}{2}(x_n - x_{n-1})$$

- h_μ: Huber loss, the smooth approximation of the ℓ_1-norm parametrized by $\mu > 0$.

  $h_\mu \in C^{1,1}_{1/\mu}(\mathbb{R}^{N-1})$.

  Closed form expression of prox_{h_μ}.

Comparisons

Proposition (see [**Briceno-Arias, Pustelnik, 2021**] for detailed references)
In the context of Problem p.16, suppose that f **is $\rho-$strongly convex**, for some $\rho \in \left]0, \alpha^{-1}\right[$, and let $\tau > 0$. Then, the following holds:

1. **Gradient descent** Suppose that $\tau \in]0, 2\beta\alpha/(\beta+\alpha)[$. Then, $\mathrm{Id} - \tau(\nabla g + \nabla f)$ is $r_G(\tau)-$Lipschitz continuous, where

$$r_G(\tau) := \max\left\{|1-\tau\rho|, |1-\tau(\beta^{-1}+\alpha^{-1})|\right\} \in]0,1[\,. \tag{1}$$

In particular, the minimum in (1) is achieved at

$$\tau^* = \frac{2}{\rho + \alpha^{-1} + \beta^{-1}}$$

and

$$r_G(\tau^*) = \frac{\alpha^{-1}+\beta^{-1}-\rho}{\alpha^{-1}+\beta^{-1}+\rho}.$$

Comparisons

Proposition (see [**Briceno-Arias, Pustelnik, 2021**] for detailed references)
In the context of Problem p.16, suppose that f **is $\rho-$strongly convex**, for some $\rho \in \left]0, \alpha^{-1}\right[$, and let $\tau > 0$. Then, the following holds:

1. **FBS** Suppose that $\tau \in]0, 2\alpha[$. Then $\mathrm{prox}_{\tau g}(\mathrm{Id} - \tau \nabla f)$ is $r_{T_1}(\tau)-$Lipschitz continuous, where

$$r_{T_1}(\tau) := \max\left\{|1 - \tau\rho|, |1 - \tau\alpha^{-1}|\right\} \in]0, 1[\,. \tag{1}$$

   In particular, the minimum in (1) is achieved at

$$\tau^* = \frac{2}{\rho + \alpha^{-1}} \quad \text{and} \quad r_{T_1}(\tau^*) = \frac{\alpha^{-1} - \rho}{\alpha^{-1} + \rho}.$$

Comparisons

Proposition (see [Briceno-Arias, Pustelnik, 2021] for detailed references)
In the context of Problem p.16, suppose that f **is $\rho-$strongly convex**, for some $\rho \in \,]0, \alpha^{-1}[$, and let $\tau > 0$. Then, the following holds:

1. **FBS** Suppose that $\tau \in]0, 2\beta]$. Then $\mathrm{prox}_{\tau f}(\mathrm{Id} - \tau\nabla g)$ is $r_{T_2}(\tau)-$Lipschitz continuous, where $r_{T_2}(\tau) := \dfrac{1}{1+\tau\rho} \in]0,1[$ .In particular, the minimum in (1) is achieved at

$$\tau^* = 2\beta \quad \text{and} \quad r_{T_2}(\tau^*) = \frac{1}{1+2\beta\rho}.$$

Proposition (see [**Briceno-Arias, Pustelnik, 2021**] for detailed references)
In the context of Problem p.16, suppose that f **is $\rho-$strongly convex**, for some $\rho \in \left]0, \alpha^{-1}\right[$, and let $\tau > 0$. Then, the following holds:

1. **PRS** $(2\mathrm{prox}_{\tau g} - \mathrm{Id}) \circ (2\mathrm{prox}_{\tau f} - \mathrm{Id})$ and $(2\mathrm{prox}_{\tau f} - \mathrm{Id}) \circ (2\mathrm{prox}_{\tau g} - \mathrm{Id})$ are $r_R(\tau)-$Lipschitz continuous, where

$$r_R(\tau) = \max\left\{\frac{1-\tau\rho}{1+\tau\rho}, \frac{\tau\alpha^{-1}-1}{\tau\alpha^{-1}+1}\right\} \in \left]0,1\right[. \tag{1}$$

In particular, the minimum in (1) is achieved at

$$\tau^* = \sqrt{\frac{\alpha}{\rho}} \quad \text{and} \quad r_R(\tau^*) = \frac{1-\sqrt{\alpha\rho}}{1+\sqrt{\alpha\rho}}.$$

Comparisons

Proposition (see [**Briceno-Arias, Pustelnik, 2021**] for detailed references)
In the context of Problem p.16, suppose that f **is $\rho-$strongly convex**, for some $\rho \in \left]0, \alpha^{-1}\right[$, and let $\tau > 0$. Then, the following holds:

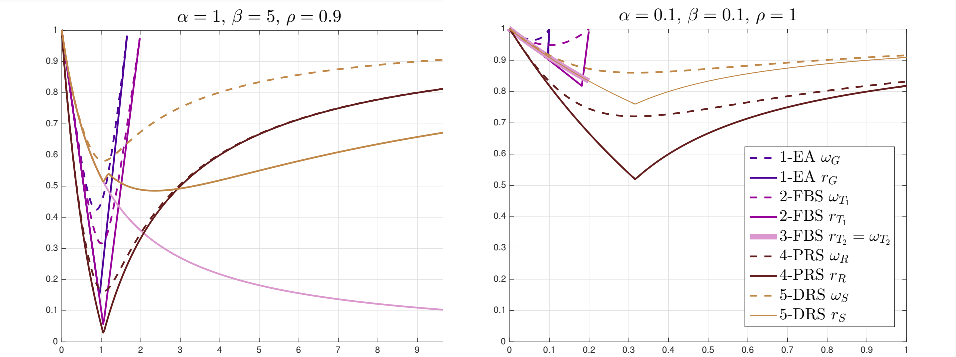
1. **DRS** $S_{\tau\nabla g,\tau\nabla f}$ and $S_{\tau\nabla f,\tau\nabla g}$ are $r_S(\tau)-$Lipschitz continuous, where

$$r_S(\tau) = \min\left\{\frac{1 + r_R(\tau)}{2}, \frac{\beta + \tau^2\rho}{\beta + \tau\beta\rho + \tau^2\rho}\right\} \in]0, 1[\tag{1}$$

and r_R is defined in p.16. In particular, the optimal step-size and the minimum in (1) are

$$(\tau^*, r_S(\tau^*)) = \begin{cases} \left(\sqrt{\frac{\alpha}{\rho}}, \frac{1}{1+\sqrt{\alpha\rho}}\right), & \text{if } \beta \leq 4\alpha; \\ \left(\sqrt{\frac{\beta}{\rho}}, \frac{2}{2+\sqrt{\beta\rho}}\right), & \text{otherwise.} \end{cases}$$

Theoretical comparisons



Comparison of the convergence rates of EA, FBS, PRS, DRS for two choices of α, β, and ρ. Note that optimization rates are better than cocoercive rates in general.

Example: Smooth TV denoising

- **First formulation**: $\underset{x\in\mathbb{R}^N}{\text{minimize}} \quad \underbrace{\frac{1}{2}\|x-z\|_2^2}_{f(x)} + \underbrace{\chi h(Lx)}_{g(x)}$

  $\rightarrow f$ is $\rho = 1$ strongly convex, $\alpha = 1$, and $\beta = \frac{\mu}{\chi\|L\|^2}$.

1- **EA:** Use $G_{\tau(\nabla g + \nabla f)}$

2- **FBS:** Use $T_{\tau\nabla f, \tau\nabla g}$

- **Second formulation**: $\min_{x\in\mathcal{H}} \underbrace{\frac{1}{2}\|x-z\|_2^2 + \chi h_{\mathbb{I}_1}(L_{\mathbb{I}_1}x)}_{\tilde{f}(x)} + \underbrace{\chi h_{\mathbb{I}_2}(L_{\mathbb{I}_2}x)}_{\tilde{g}(x)}$

  $\rightarrow \widetilde{f}$ is $\rho = 1$ strongly convex, $\alpha = \frac{\mu}{\mu+\chi\|L_{\mathbb{I}_2}\|^2}$, and $\beta = \frac{\mu}{\chi\|L_{\mathbb{I}_1}\|^2}$.

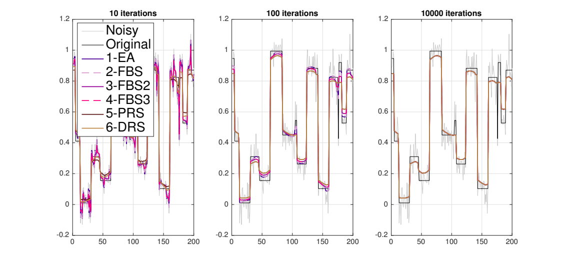
3- **FBS 2:** Use $T_{\tau\nabla\widetilde{g}, \tau\nabla\widetilde{f}}$

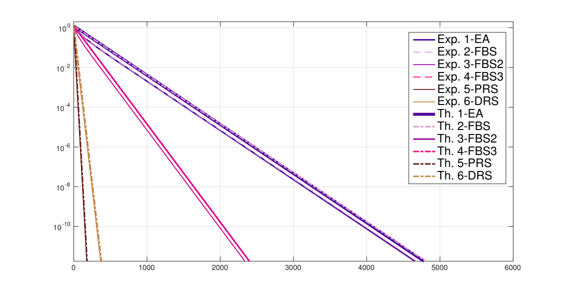
4- **FBS 3:** Use $T_{\tau\nabla\widetilde{f}, \tau\nabla\widetilde{g}}$

5- **PRS:** Use $R_{\tau\nabla\widetilde{f}, \tau\nabla\widetilde{g}}$

6- **DRS:** Use $S_{\tau\nabla\widetilde{f}, \tau\nabla\widetilde{g}}$

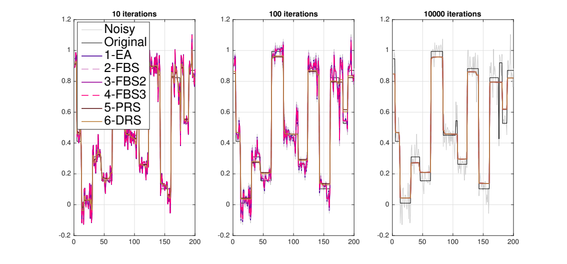
Numerical and theoretical comparisons

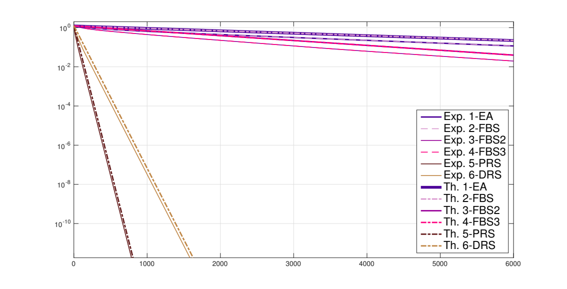




Piecewise constant denoising estimates after 10, 100, and 10000 iterations with $\chi = 0.7$ and $\mu = 0.002$ when considering gradient descent, FBS, PRS or DRS. Associated theoretical and numerical convergence rates.

Numerical and theoretical comparisons





Piecewise constant denoising estimates after 10, 100, and 10000 iterations with $\chi = 0.7$ and $\mu = 0.0001$ when considering gradient descent, FBS, PRS or DRS. Associated theoretical and numerical convergence rates.

Optimization algorithms

Forward-Backward	$f_1 + f_2$	$f_1 \in C^{1,1}_{\zeta}(\mathcal{H})$ prox_{f_2}	[Combettes,Wajs,2005]
ISTA	$f_1 + f_2$	$f_1 \in C^{1,1}_{\zeta}(\mathcal{H})$ $f_2 = \lambda\|\cdot\|_1$	[Daubechies et al, 2003]
Projected gradient	$f_1 + f_2$	$f_1 \in C^{1,1}_{\zeta}(\mathcal{H})$ $f_2 = \iota_C$	
Gradient descent	$f_1 + f_2$	$f_1 \in C^{1,1}_{\zeta}(\mathcal{H})$ $f_2 = 0$	
Douglas-Rachford	$f_1 + f_2$	prox_{f_1} prox_{f_2}	[Combettes,Pesquet, 2007]
PPXA	$\sum_i f_i$	prox_{f_i}	[Combettes,Pesquet, 2008]