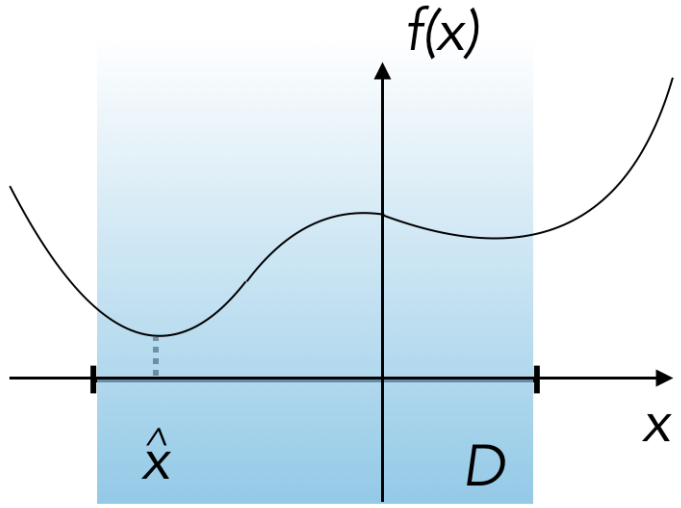
Mathematical foundations in deep learning
Part I: Optimization – Introduction

Nelly Pustelnik

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

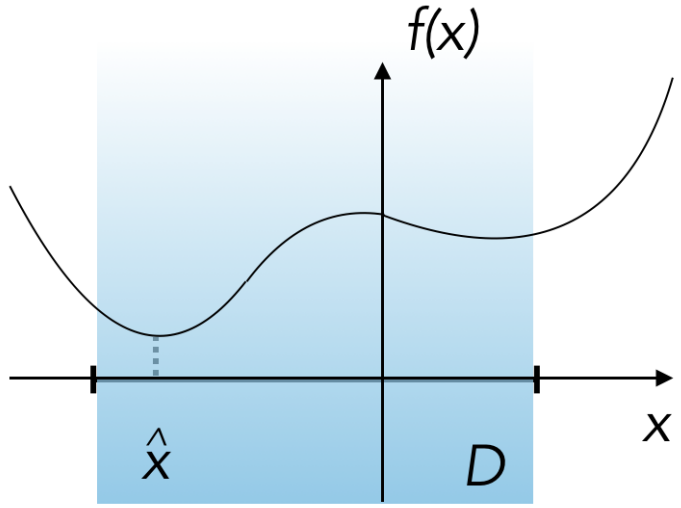
Minimization problems

- **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 $\mathbb{R}^N$.

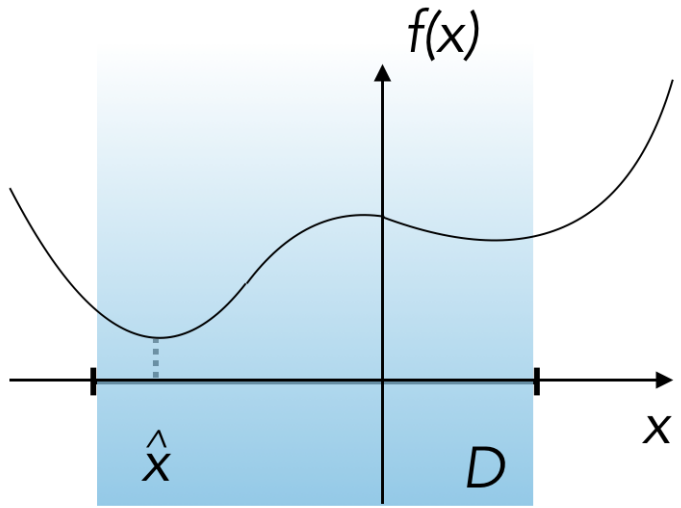


- Goal: We want to

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

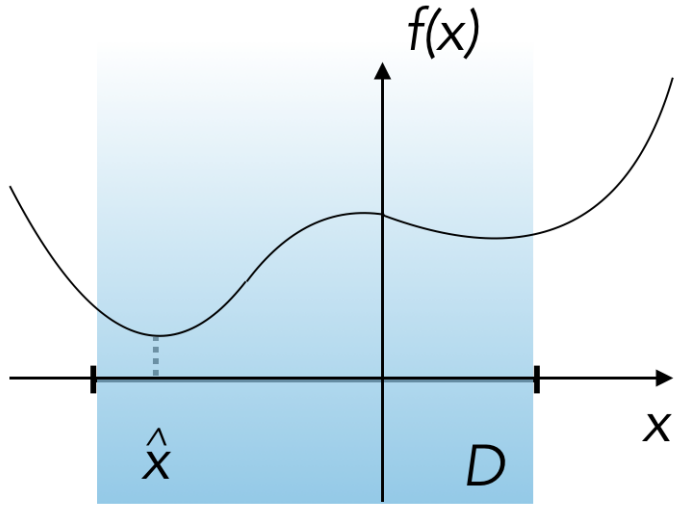
Minimization problems

- **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 $\mathbb{R}^N$.



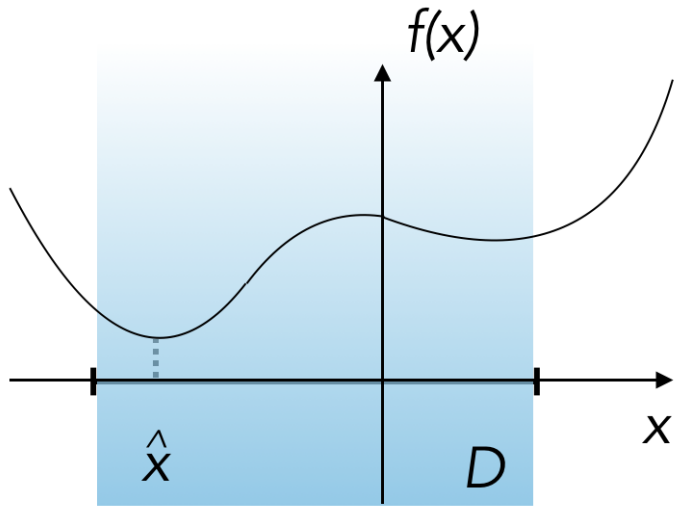
- Goal: We want to

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

Minimization problems

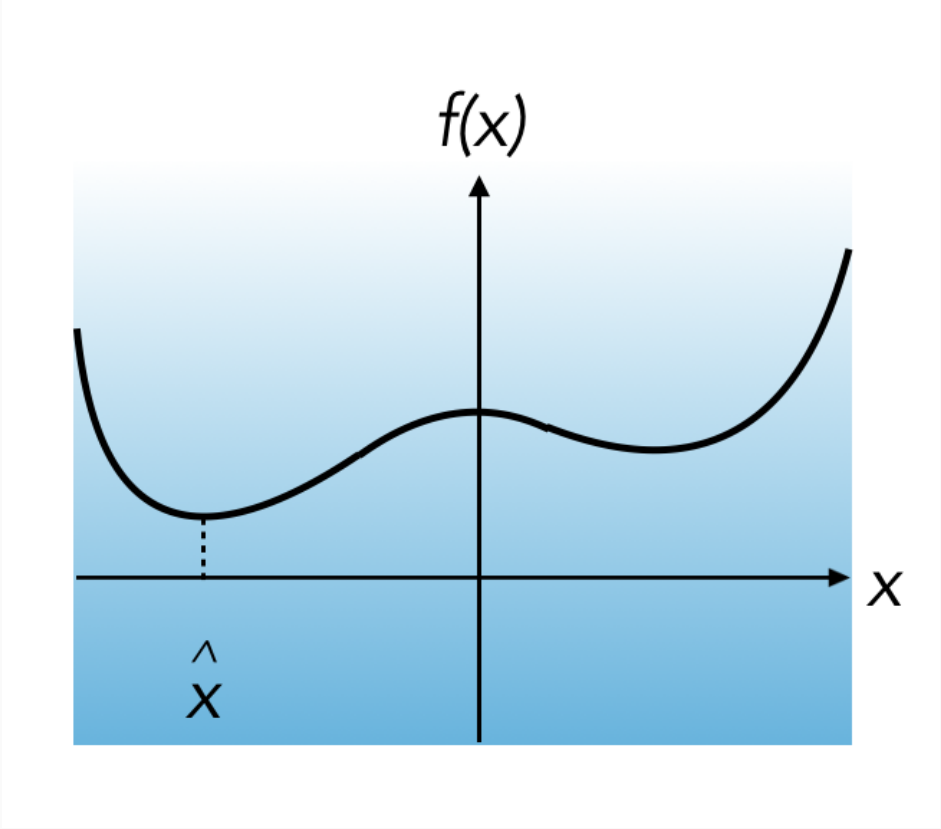
- **Maximization problems** involving :
  - a reward function $f : \mathbb{R}^N \to \mathbb{R}$;
  - a subset D of $\mathbb{R}^N$.

Without loss of generality, we can focus on minimization problems



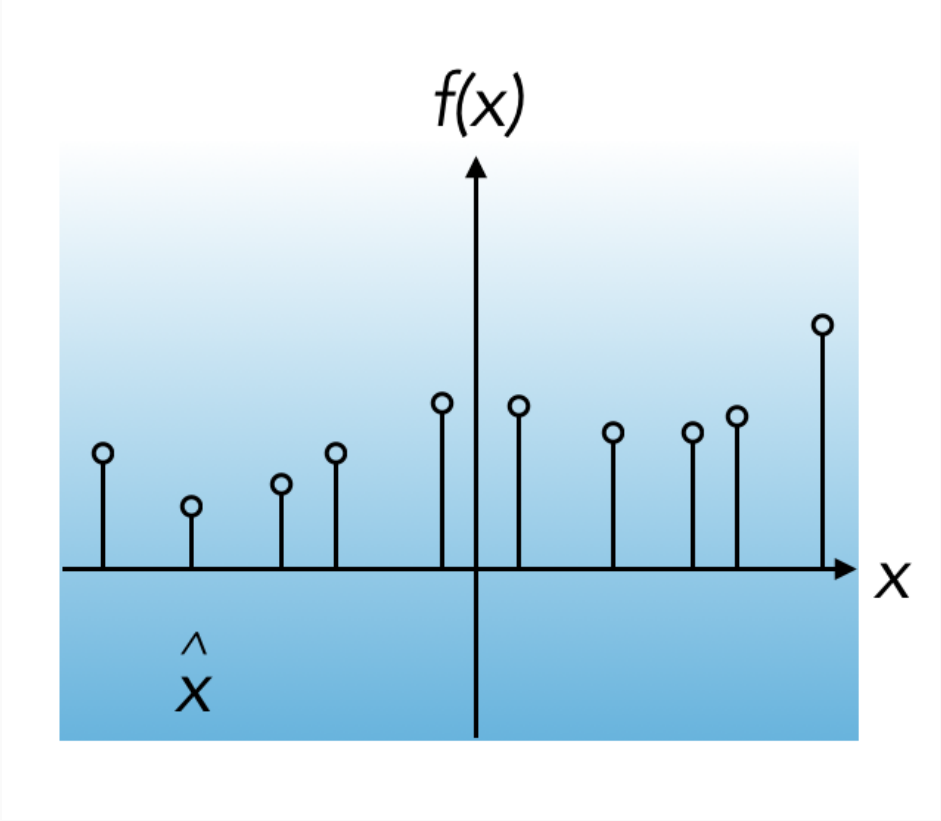
Various types of minimization problems

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



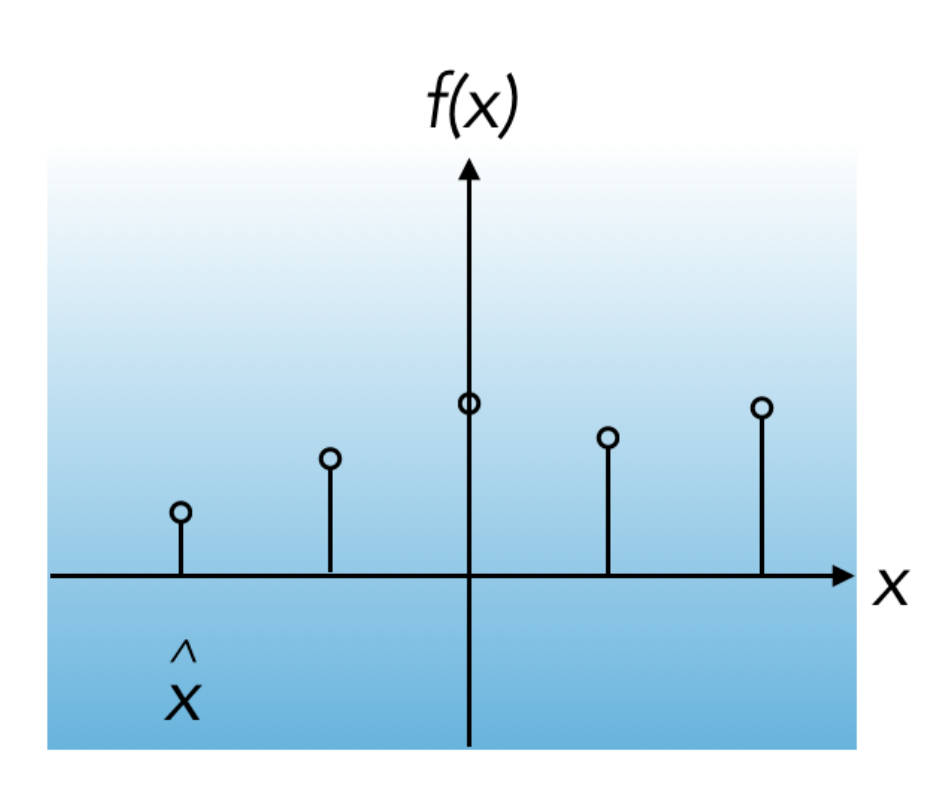
Various types of minimization problems

- 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

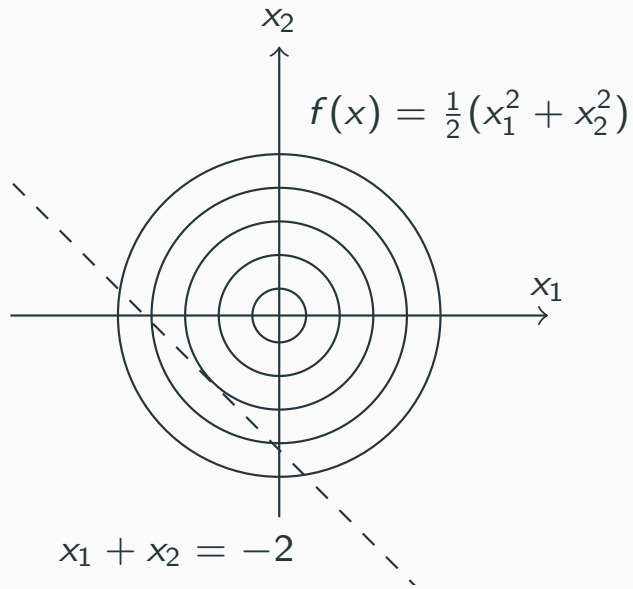
    $$D = \left\{ x \in \mathbb{R}^N \mid (\forall i \in \{1, \dots, P\}) \quad g_i(x) = 0 \right\}$$

    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 $b_i \in \mathbb{R}$.



Various types of minimization problems

- D **uncountable**: continuous optimization problem

  - Example: Optimization problem with P inequality constraints

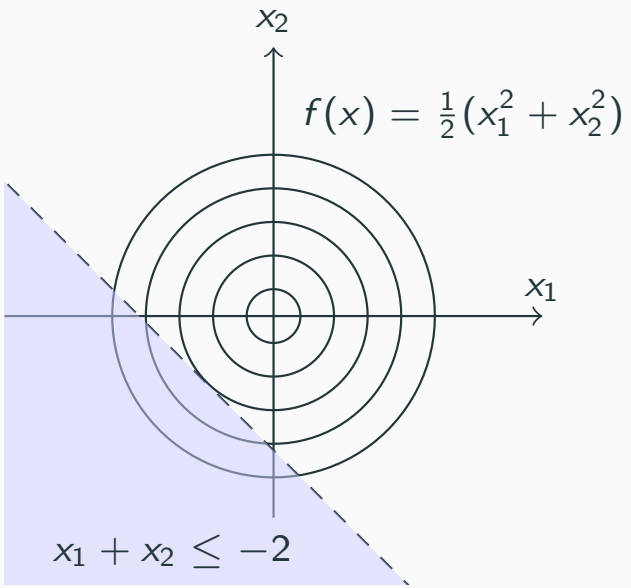
    $$D = \left\{ x \in \mathbb{R}^N \mid (\forall i \in \{1, \dots, P\}) \quad g_i(x) \leq 0 \right\}$$

    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 $\mathrm{b}_i \in \mathbb{R}$.



Constrained and unconstrained minimization problems

- Reformulation using **indicator function**

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

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**

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

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

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

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**

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

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

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 $d : \mathcal{H} \to \mathcal{G}$




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 $d : \mathcal{H} \to \mathcal{G}$






Example: Supervised learning

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

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

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

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

Example: Supervised learning

- **Database**: $\mathcal{S} = \left\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\middle|\; \ell \in \{1, \dots, L\}\right\}$
- **Goal**: Learn a prediction function $d : \mathcal{H} \to \mathcal{G}$
- 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: $\underset{x \in \mathcal{H}}{\text{minimize}} \;\; \frac{1}{L}\sum_{\ell=1}^{L} \log\left(1 + e^{-z_\ell x^\top u_\ell}\right) + \lambda \|x\|_2^2$

Example: Supervised learning

- **Database**: $\mathcal{S} = \left\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\middle|\; \ell \in \{1, \dots, L\}\right\}$
- **Goal**: Learn a prediction function $d : \mathcal{H} \to \mathcal{G}$
- 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: $\underset{x \in \mathcal{H}}{\text{minimize}} \; \frac{1}{L}\sum_{\ell=1}^{L}\log\left(1 + e^{-z_\ell x^\top u_\ell}\right) + \lambda\|x\|_2^2$

⇒ **Convex smooth problems**

Example: Supervised learning

- **Database**: $\mathcal{S} = \big\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\big|\; \ell \in \{1, \dots, L\}\big\}$
- **Goal**: Learn a prediction function $d : \mathcal{H} \to \mathcal{G}$
- 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)

  - $\odot$ Ridge regression: $\underset{x \in \mathcal{H}}{\text{minimize}} \;\; \frac{1}{L}\sum_{\ell=1}^{L}(z_\ell - x^\top u_\ell)^2 + \lambda \|x\|_2^2$
  - $\odot$ 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**: $\mathcal{S} = \left\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\middle|\; \ell \in \{1, \dots, L\}\right\}$
- **Goal**: Learn a prediction function $d : \mathcal{H} \to \mathcal{G}$
- 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: $\underset{x \in \mathcal{H}}{\text{minimize}} \;\; \frac{1}{L}\sum_{\ell=1}^{L} \log\left(1 + e^{-z_\ell x^\top u_\ell}\right) + \lambda \|x\|_1$
  - ⊙ SVM classification: $\underset{x \in \mathcal{H}}{\text{minimize}} \;\; \frac{1}{L}\sum_{\ell=1}^{L} \max\left(0, 1 - z_\ell x^\top u_\ell\right) + \lambda \|x\|_2^2$

Example: Supervised learning

- **Database**: $\mathcal{S} = \left\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\middle|\; \ell \in \{1, \dots, L\}\right\}$
- **Goal**: Learn a prediction function $d : \mathcal{H} \to \mathcal{G}$
- 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: $\underset{x \in \mathcal{H}}{\text{minimize}} \;\; \frac{1}{L}\sum_{\ell=1}^{L}\log\left(1 + e^{-z_\ell x^\top u_\ell}\right) + \lambda\|x\|_1$
  - ⊙ SVM classification: $\underset{x \in \mathcal{H}}{\text{minimize}} \;\; \frac{1}{L}\sum_{\ell=1}^{L}\max\left(0, 1 - z_\ell x^\top u_\ell\right) + \lambda\|x\|_2^2$

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

Example: Supervised learning

- **Database**: $\mathcal{S} = \left\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\middle|\; \ell \in \{1, \dots, L\}\right\}$
- **Goal**: Learn a prediction function $d : \mathcal{H} \to \mathcal{G}$
- 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):

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

  - ⊙ 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**: $\mathcal{S} = \left\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\middle|\; \ell \in \{1, \dots, L\}\right\}$
- **Goal**: Learn a prediction function $d : \mathcal{H} \to \mathcal{G}$
- 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):

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

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

Example: Supervised learning

- **Database**: $\mathcal{S} = \left\{(u_\ell, z_\ell) \in \mathcal{H} \times \mathcal{G} \;\middle|\; \ell \in \{1, \dots, L\}\right\}$
- **Goal**: Learn a prediction function $d : \mathcal{H} \to \mathcal{G}$
- 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):

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

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

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

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 $]-\infty,+\infty]$:

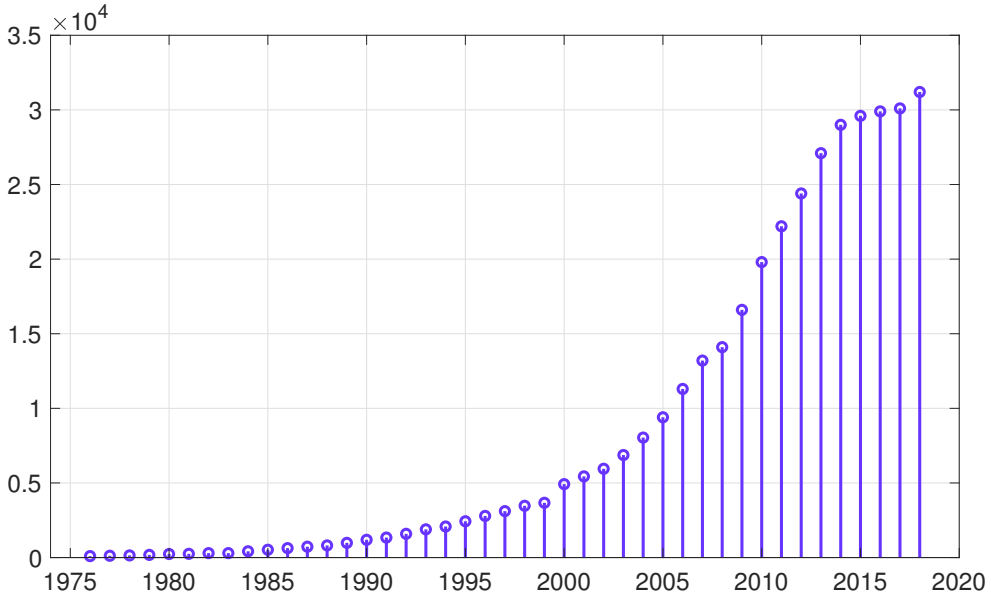
$$\widehat{x} \in \underset{x \in \mathbb{R}^N}{\operatorname{Argmin}} \sum_{s=1}^{S} f_s(L_s x)$$

- **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

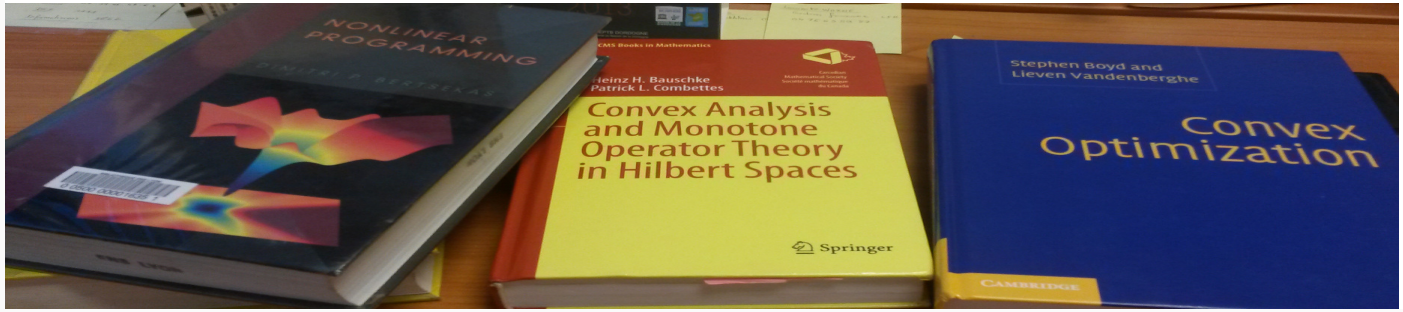
General objective function involving linear operators L_s from $\mathbb{R}^N$ to $\mathbb{R}^{|\Upsilon_s|}$ and functions f_s proper, convex, l.s.c functions from $\mathbb{R}^{|\Upsilon_s|}$ to $]-\infty,+\infty]$:

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



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]

Mathematical foundations in deep learning
Part II: Optimization – Basics

Nelly Pustelnik

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

(several slides in this part traced back to 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:

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

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

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

Norm and adjoint

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:

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

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

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

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

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

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

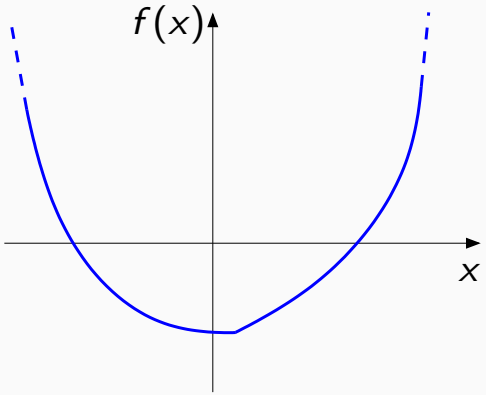
- Proper function
- Lower semi-continuous function
- Convex function

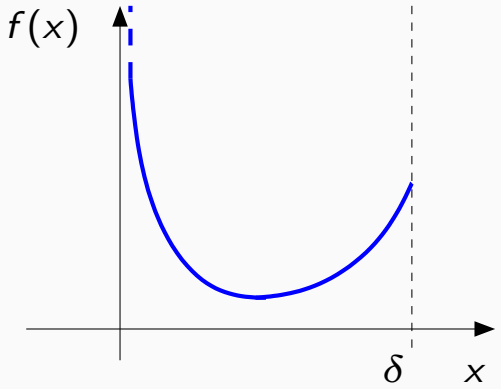
Functional analysis: definitions

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

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

Domains of the functions ?



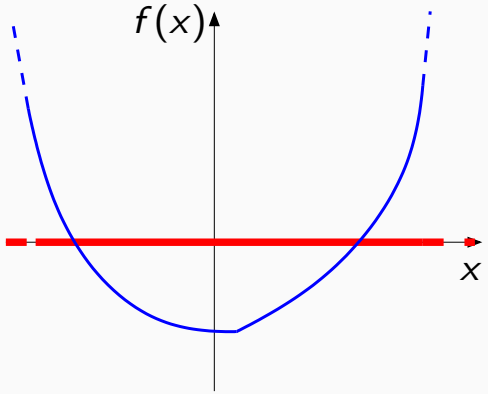


Functional analysis: definitions

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

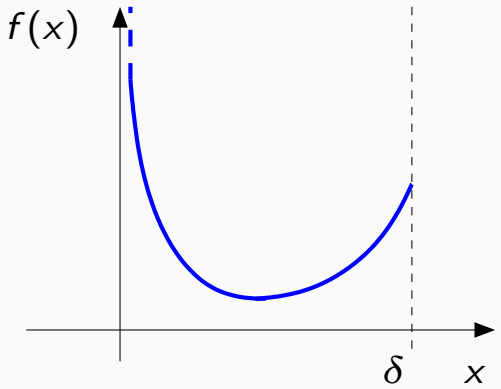
- The **domain** of f is $\operatorname{dom} f = \{x \in \mathcal{H} \mid f(x) < +\infty\}$.
- The function f is **proper** if $\operatorname{dom} f \neq \varnothing$.

Domains of the functions ?



$\operatorname{dom} f = \mathbb{R}$

(proper)

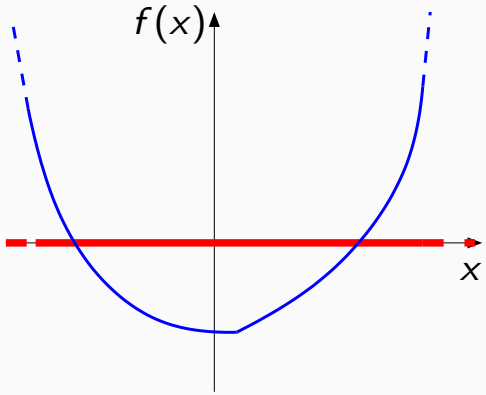


Functional analysis: definitions

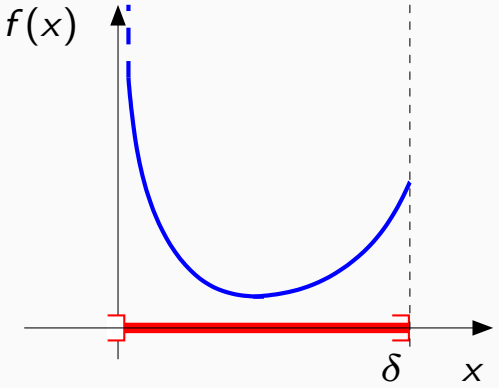
Let $f : \mathcal{H} \to]-\infty, +\infty]$ where $\mathcal{H}$ is a Hilbert space.

- 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)

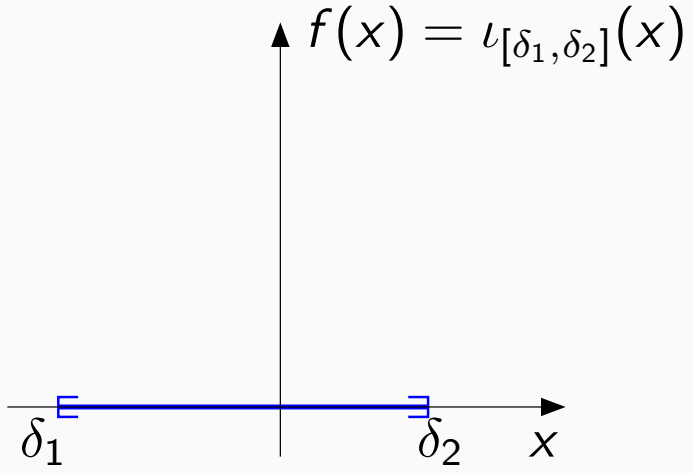
Functional analysis: definitions

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

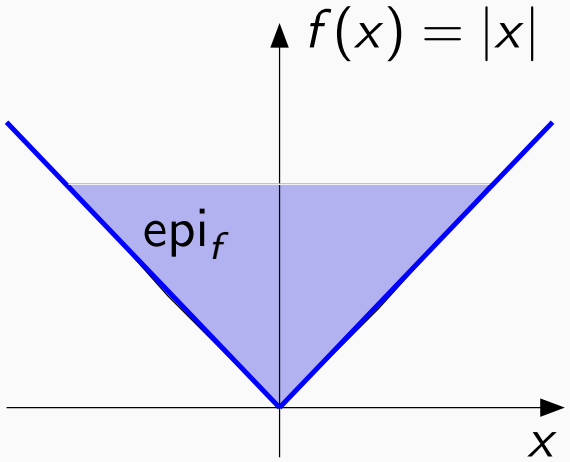
Let $f : \mathcal{H} \to]-\infty, +\infty]$. The **epigraph** of f is

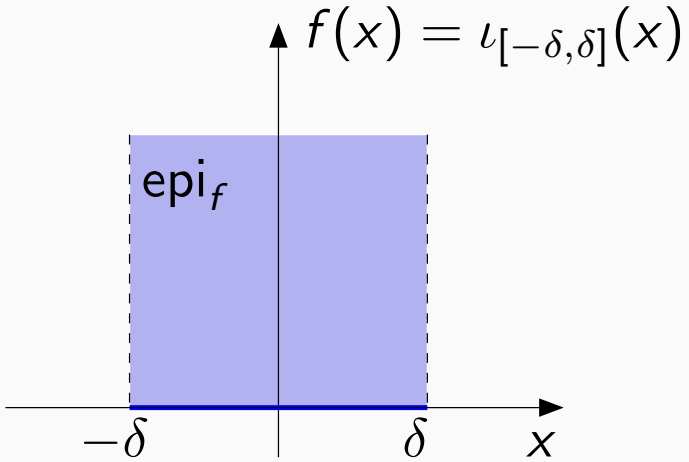
$$\operatorname{epi} f = \left\{(x, \zeta) \in \operatorname{dom} f \times \mathbb{R} \;\middle|\; f(x) \leq \zeta\right\}$$

Epigraph

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

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

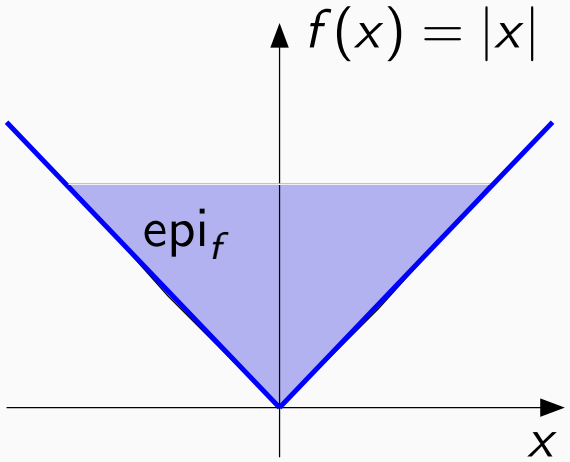


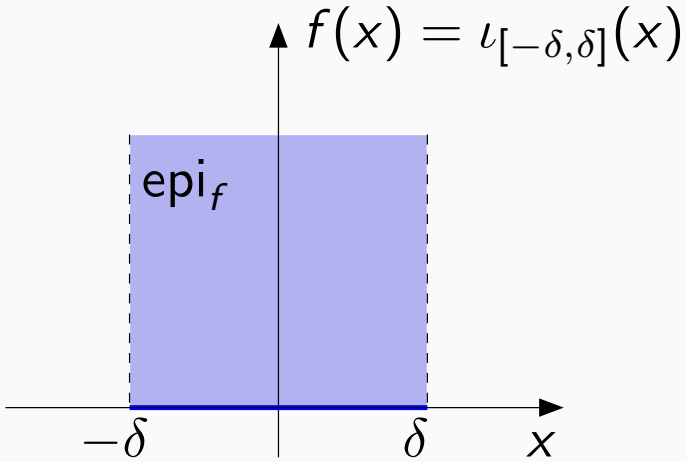


Epigraph

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

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





- 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**: $\sum_n |x_n| \leq \eta \Leftrightarrow \begin{cases} |x_n| \leq \zeta_n \\ \sum_n \zeta_n \leq \eta \end{cases}$

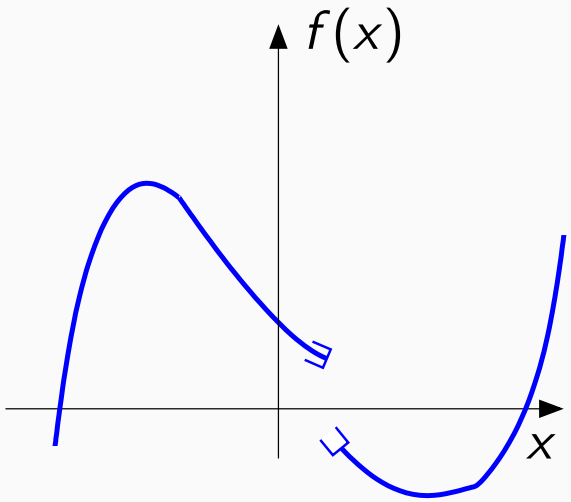
Lower semi-continuity

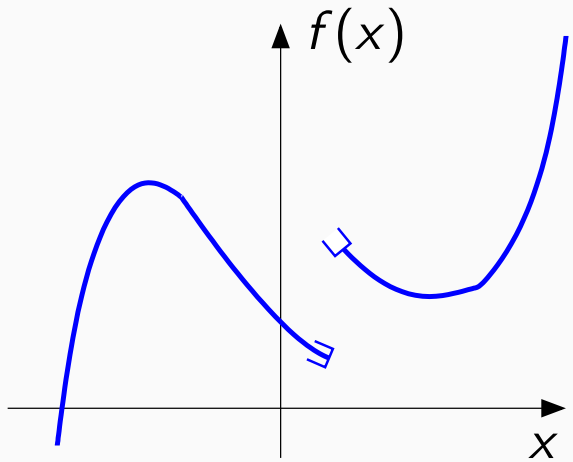
Let $f : \mathcal{H} \to]-\infty, +\infty]$.
f is a **lower semi-continuous** function on $\mathcal{H}$ if and only if epi f is closed

Lower semi-continuity

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

- l.s.c. functions ?

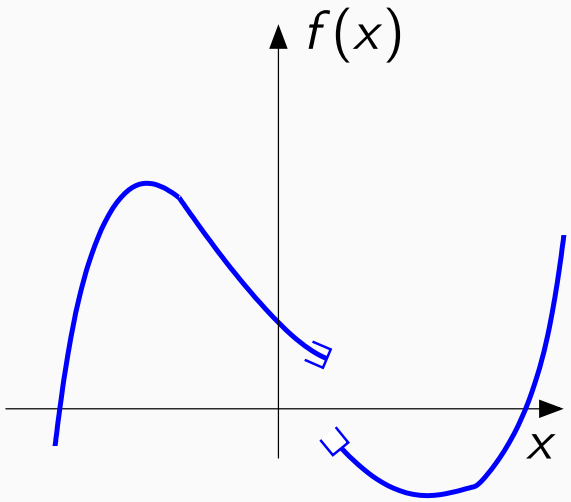


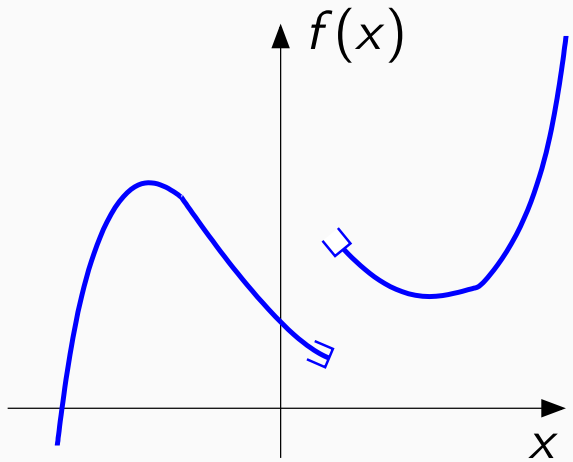


Lower semi-continuity

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 ?

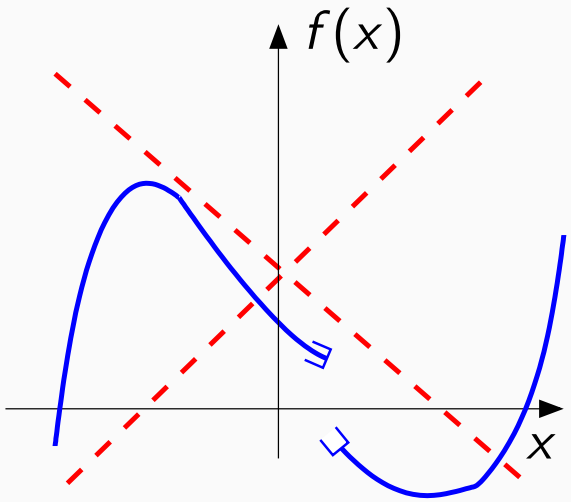


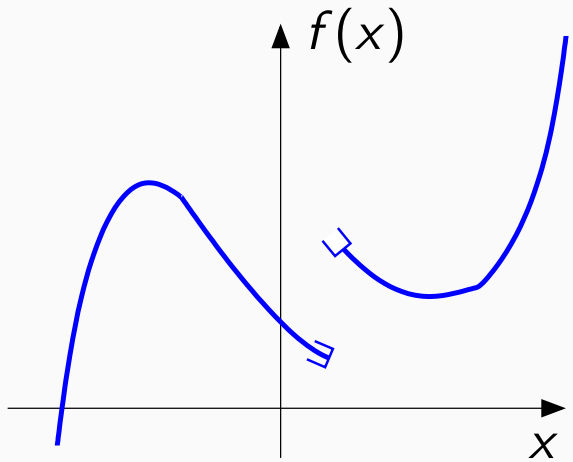


Lower semi-continuity

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

- l.s.c. functions ?

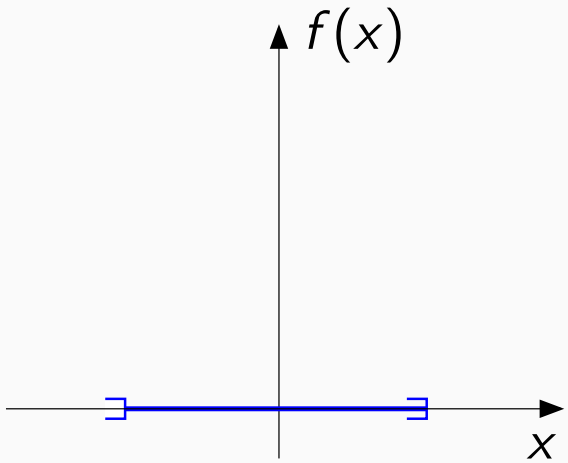


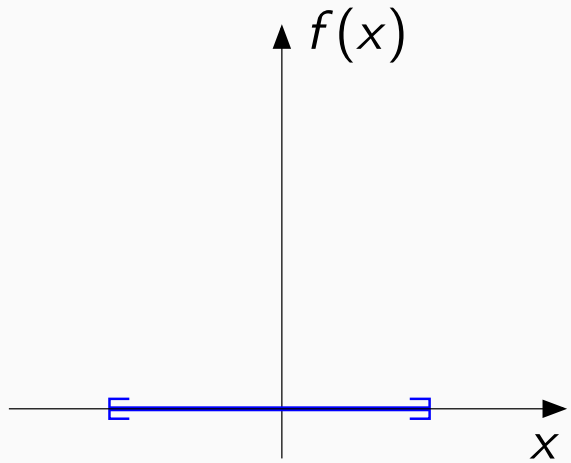


Lower semi-continuity

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 ?

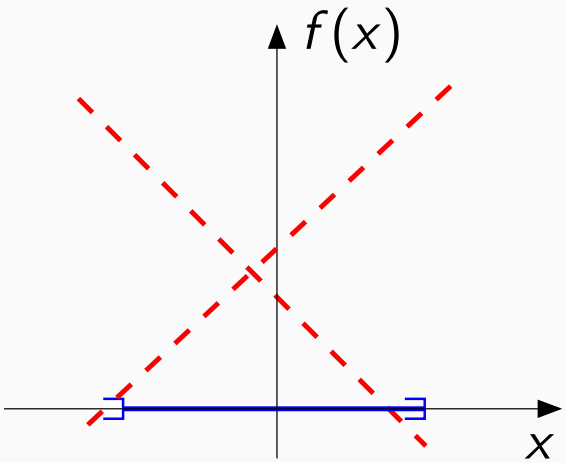


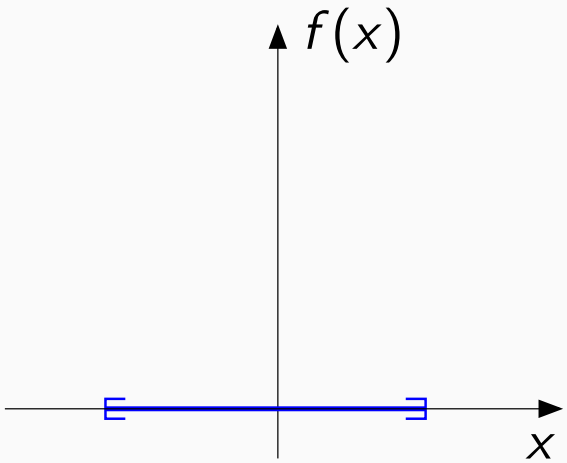


Lower semi-continuity

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

- l.s.c. functions ?

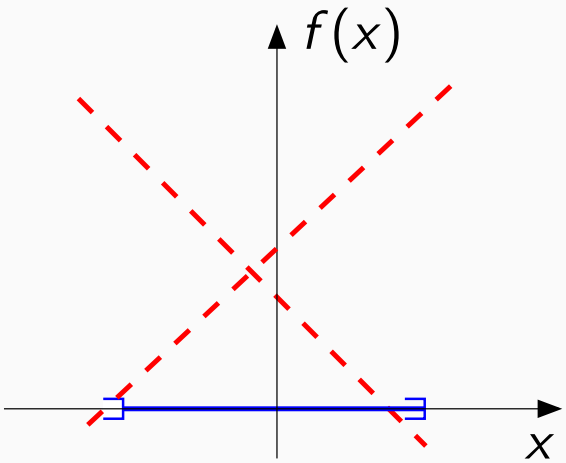


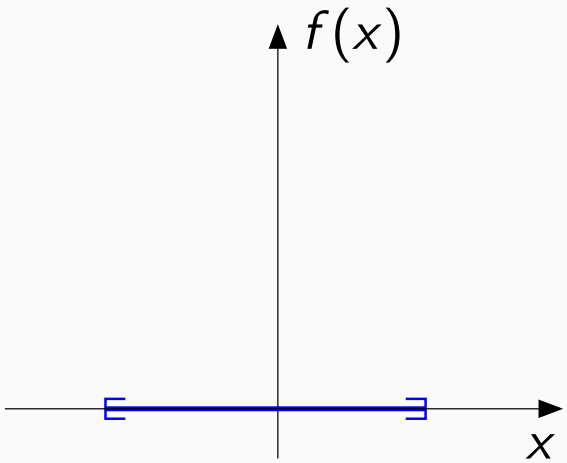


Lower semi-continuity

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 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 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 $x > 0$;

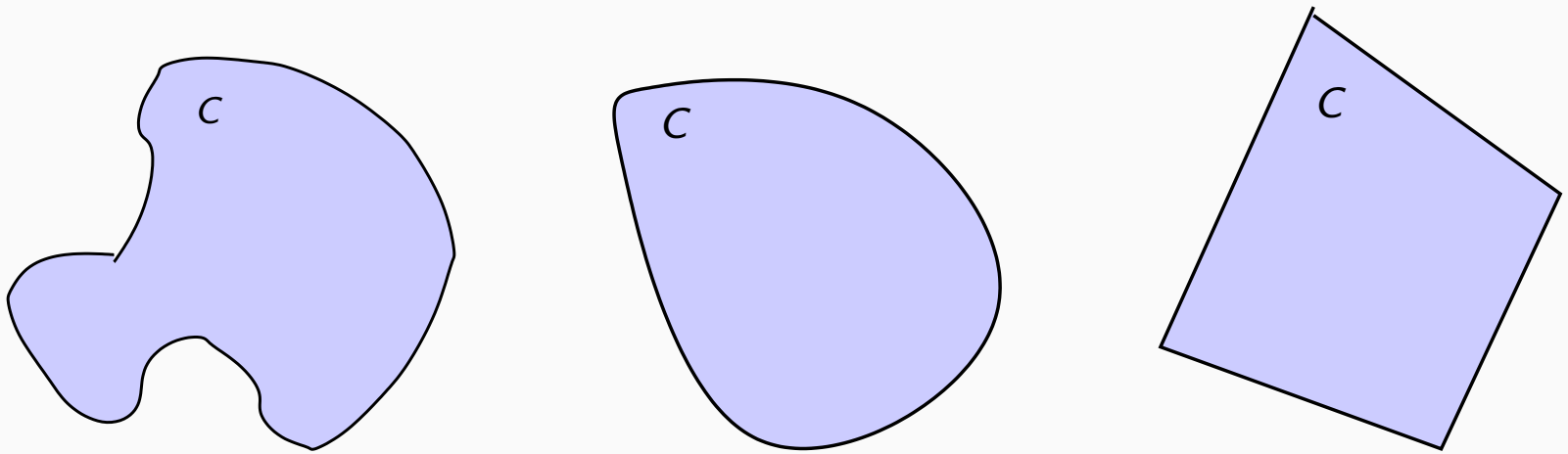
- Properties:
  - ⊙ Every continuous function on $\mathcal{H}$ is l.s.c.
  - ⊙ **Every finite sum of l.s.c. functions is l.s.c.**
  - ⊙ Let $(f_i)_{i \in I}$ be a family of l.s.c functions. Then, $\sup_{i \in I} f_i$ is l.s.c.

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 ?

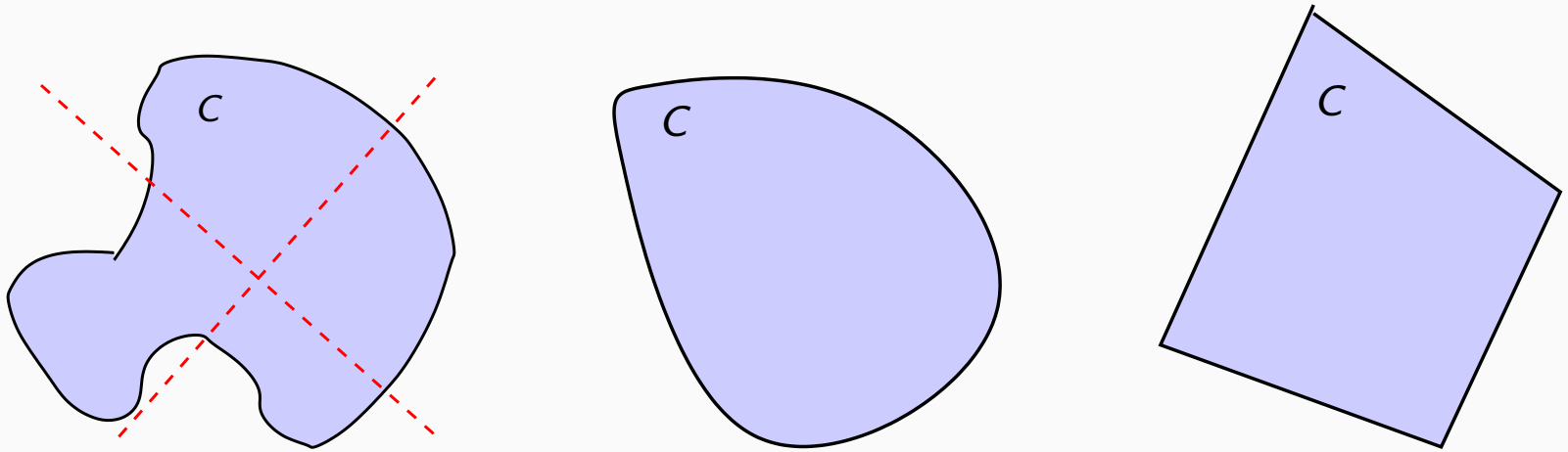


Convex set

$C \subset \mathcal{H}$ 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

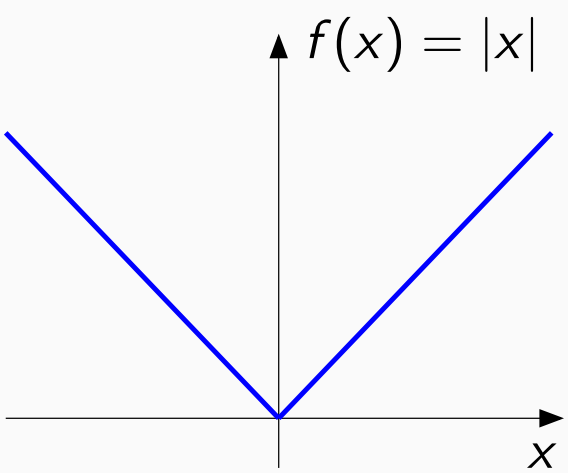
$$\big(\forall (x, y) \in \mathcal{H}^2\big)(\forall \alpha \in]0, 1[) \qquad f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

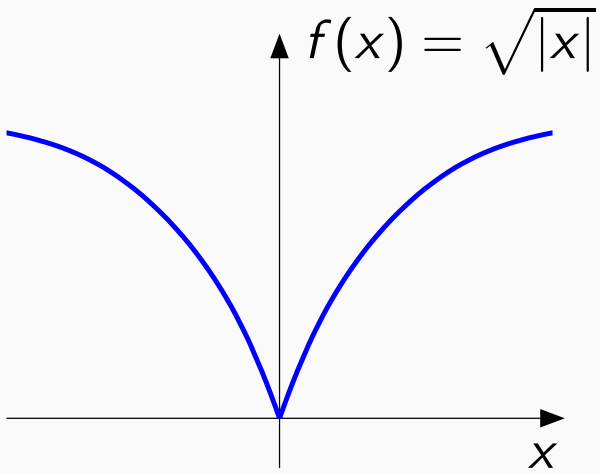
Convex function: definitions

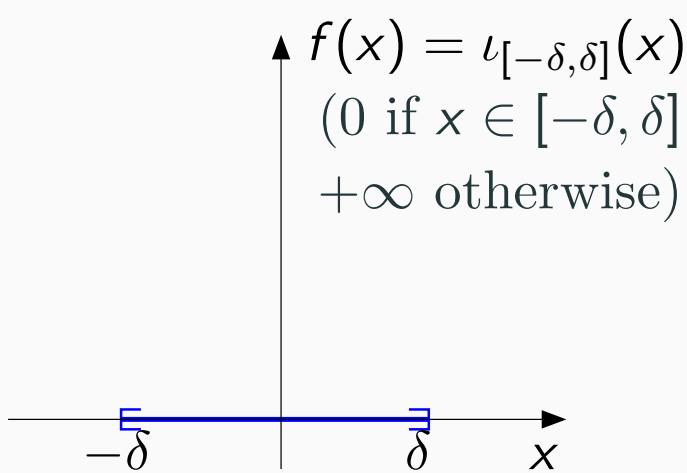
$f : \mathcal{H} \to]-\infty, +\infty]$ is a **convex function** if

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

Convex functions ?





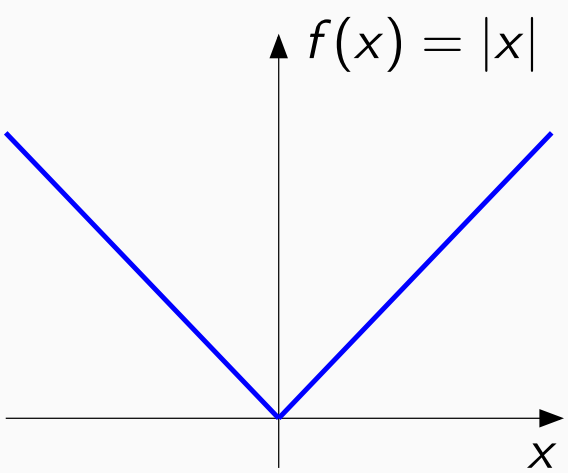


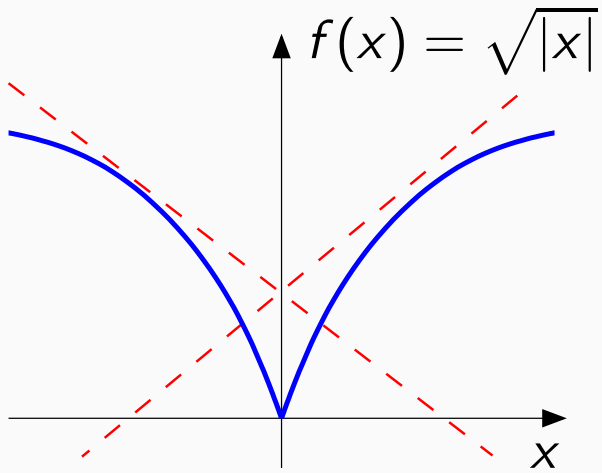
Convex function: definitions

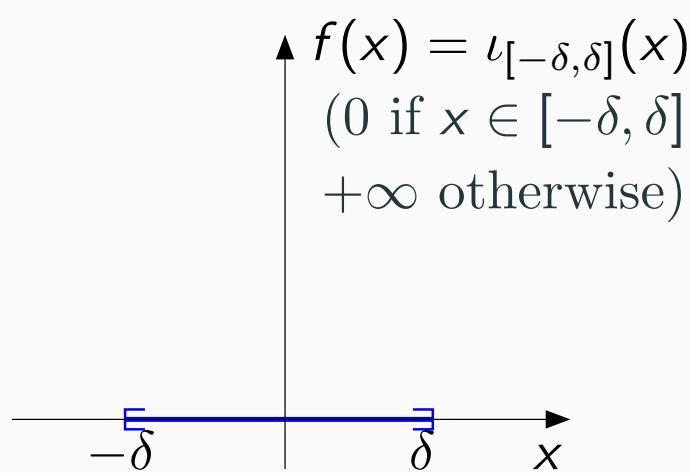
$f : \mathcal{H} \to]-\infty, +\infty]$ is a **convex function** if

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

Convex functions ?





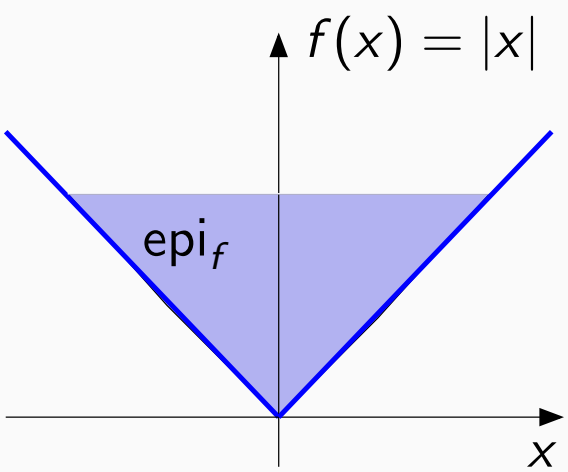


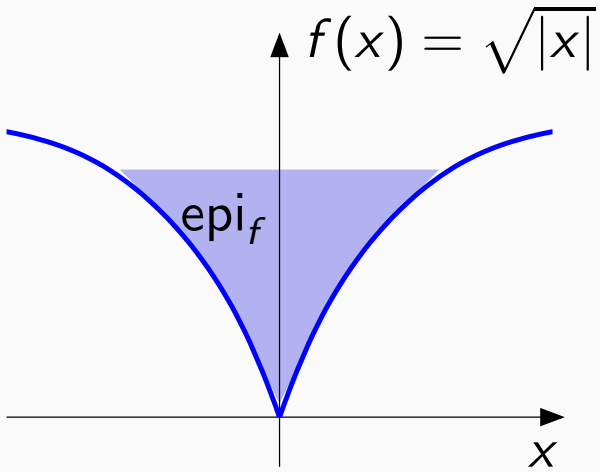
Convex functions: definition

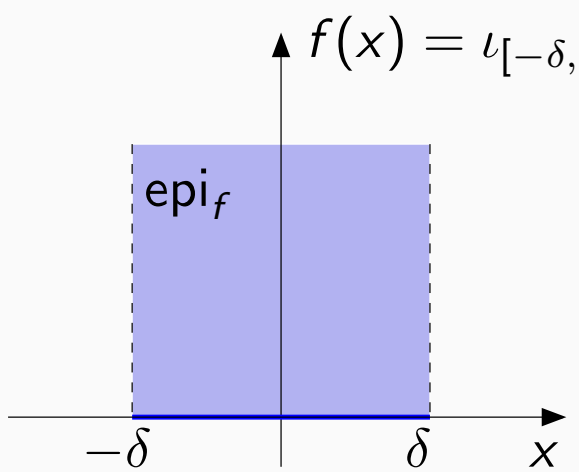
$f : \mathcal{H} \to]-\infty, +\infty]$ is convex $\Leftrightarrow$ its epigraph is convex.

Convex functions: definition

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

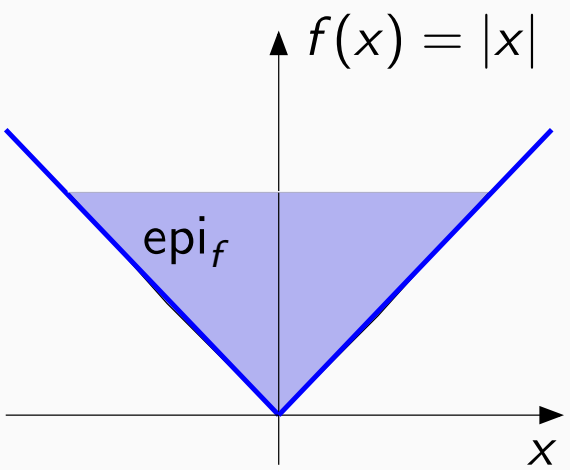


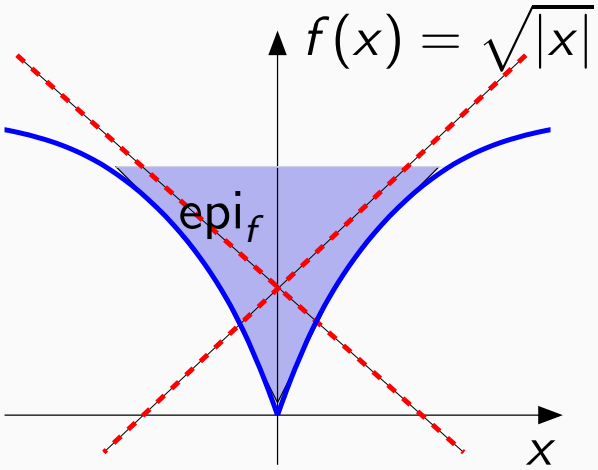


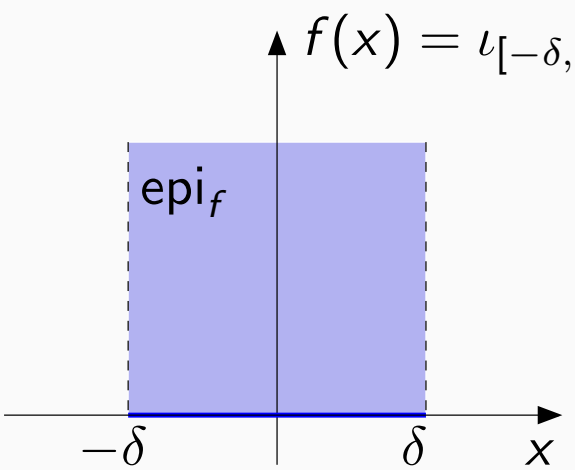


Convex functions: definition

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

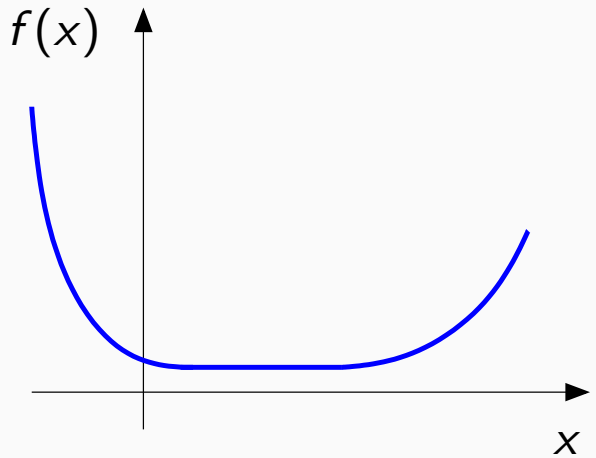
$$(\forall x \in \operatorname{dom} f)(\forall y \in \operatorname{dom} f)(\forall \alpha \in]0,1[)$$
$$x \neq y \quad \Rightarrow \quad f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y).$$

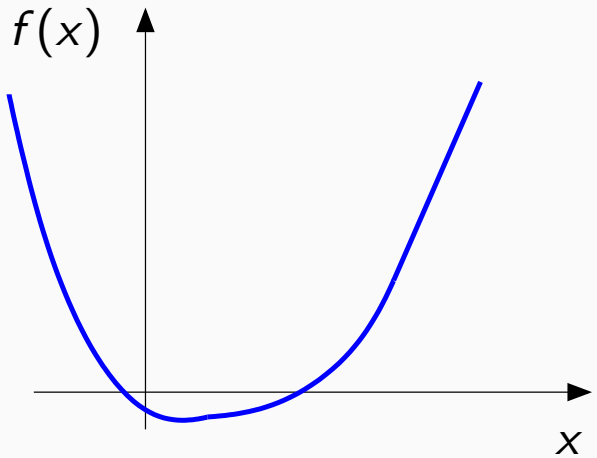
Strictly convex functions

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



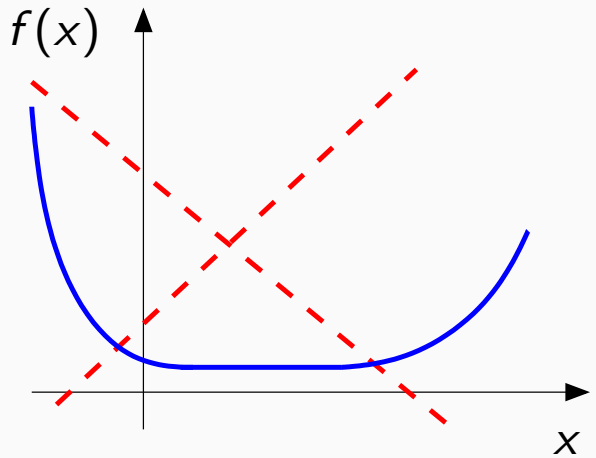


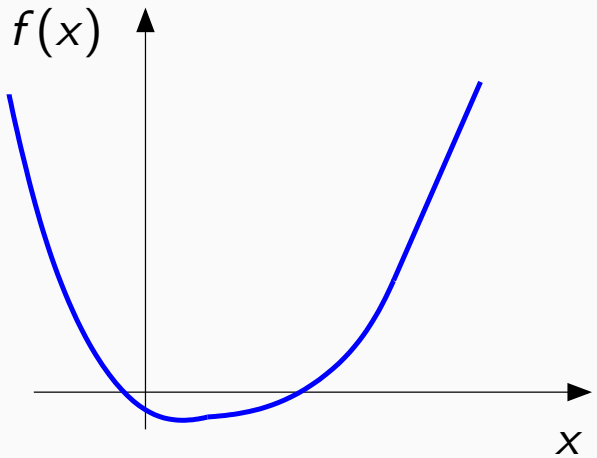
Strictly convex functions

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

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

- 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 $\mu = \inf f > -\infty$.

- $\left\{x \in \mathcal{H} \;\middle|\; f(x) = \mu\right\}$ 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 $f : \mathcal{H} \to]-\infty, +\infty]$ be a **proper convex** function such that $\mu = \inf f > -\infty$.

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

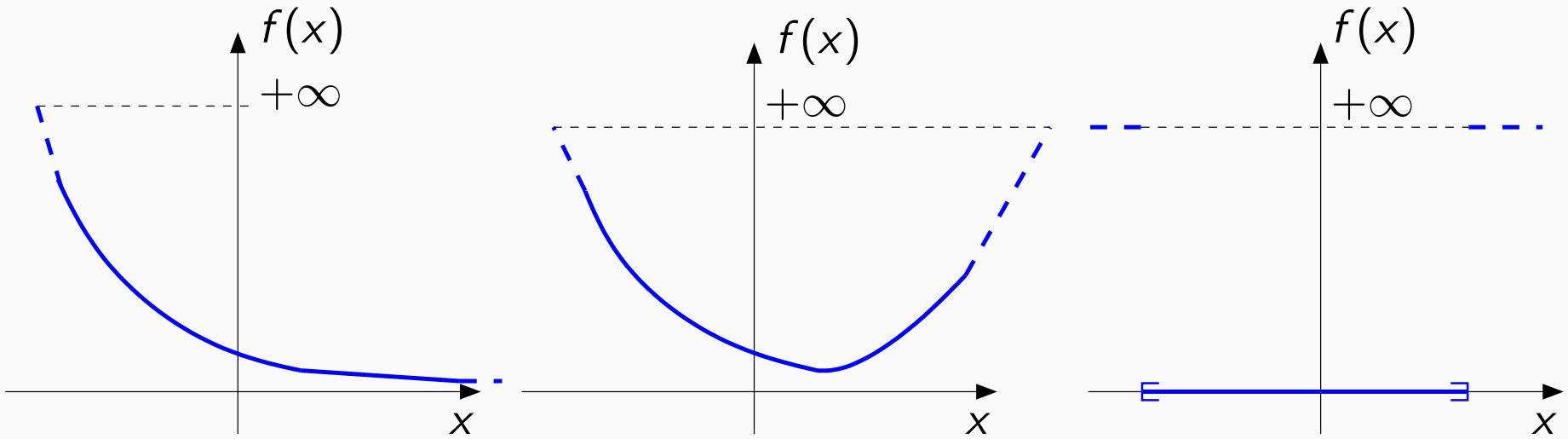
Existence of a minimizer

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

Existence of a minimizer

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

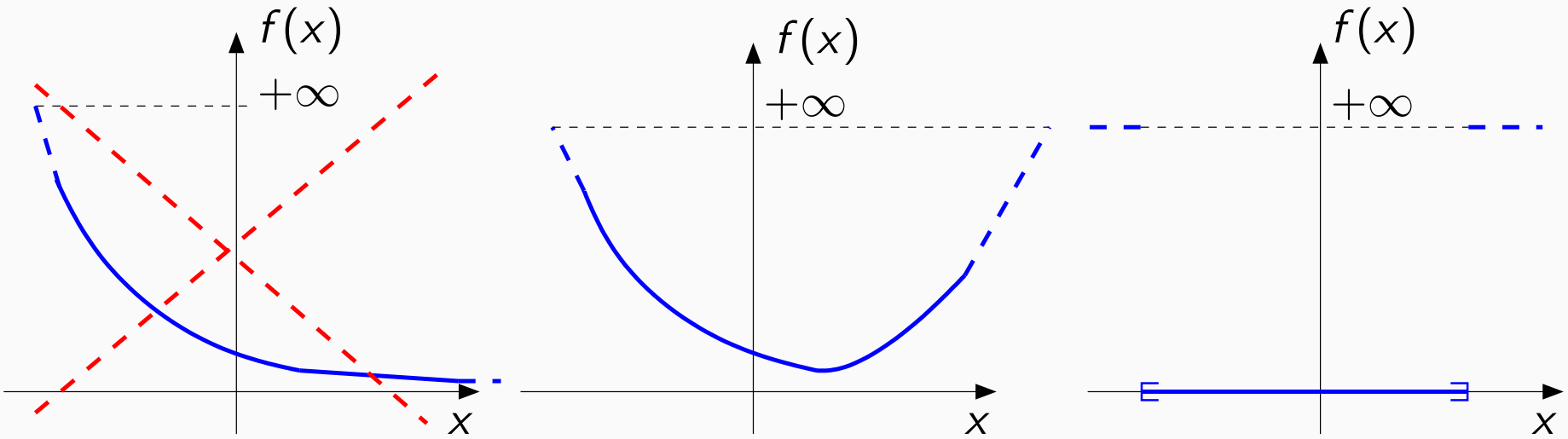
Coercive functions ?



Existence of a minimizer

Let $\mathcal{H}$ be a Hilbert space. Let $f : \mathcal{H} \to]-\infty, +\infty]$.
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 $\operatorname{dom} f \cap C \neq \varnothing$.

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

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

- 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 $f : \mathcal{H} \to]-\infty, +\infty]$ 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**

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

$\rightarrow$ Optimality condition:

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

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

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

Optimality condition

- Example: **Logistic based criterion:**

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

$\rightarrow$ Optimality condition:

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

$\rightarrow$ **No closed form expression**. An iterative procedure is required.

Iterative scheme

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

- 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 $\sup_{n \in \mathbb{N}} \gamma_n < 2\alpha^{-1}$.

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].

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 $\sum_{n=0}^{+\infty} \gamma_n = +\infty$.

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**.

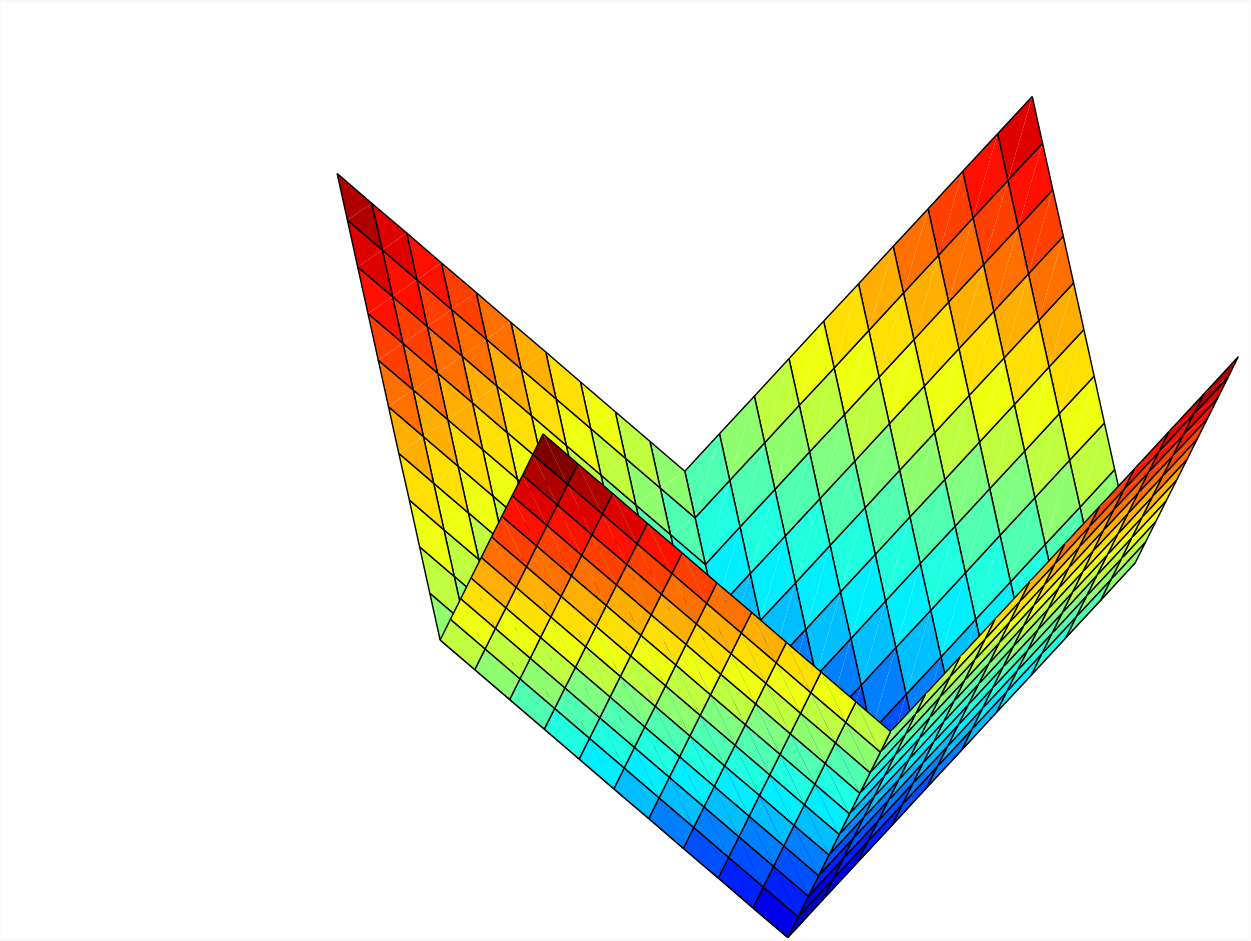
Mathematical foundations in deep learning
Part III: Optimization – Subdifferential

Nelly Pustelnik

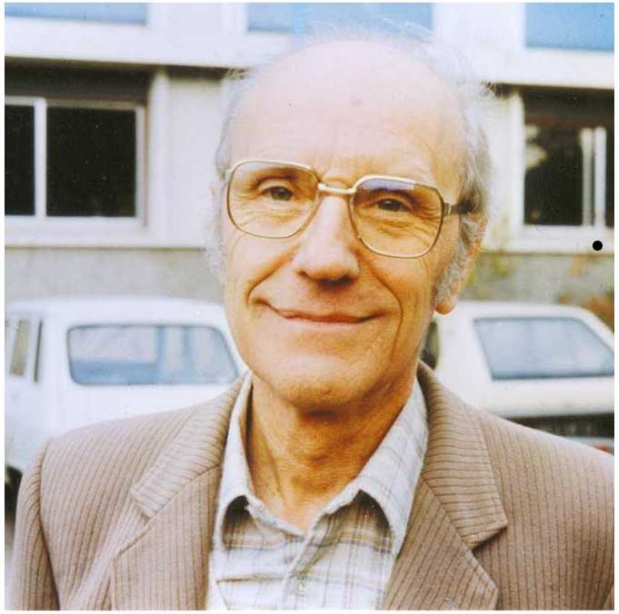
CNRS, Laboratoire de Physique de l'ENS de Lyon, France

(several slides in this part traced back to 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



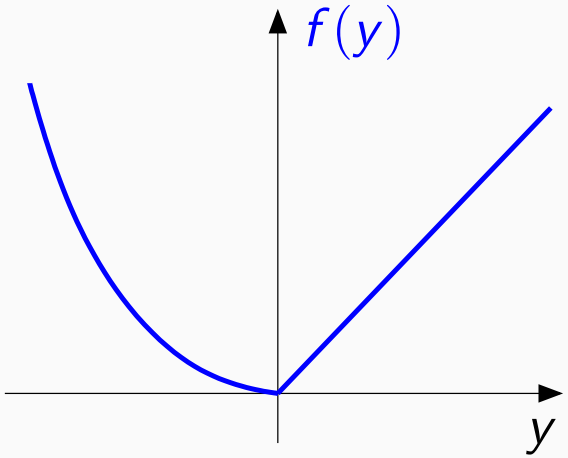
Jean-Jacques Moreau
(1923–2014)

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,

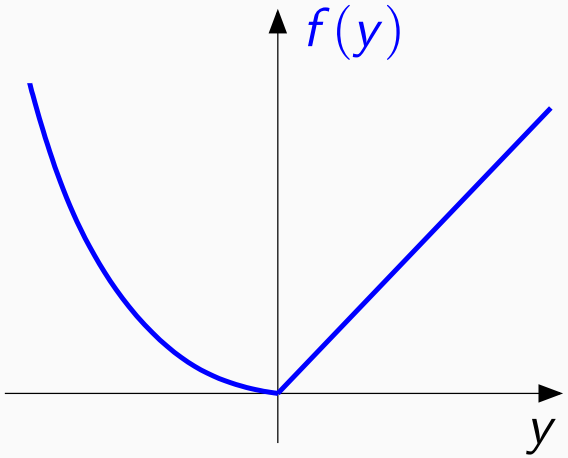


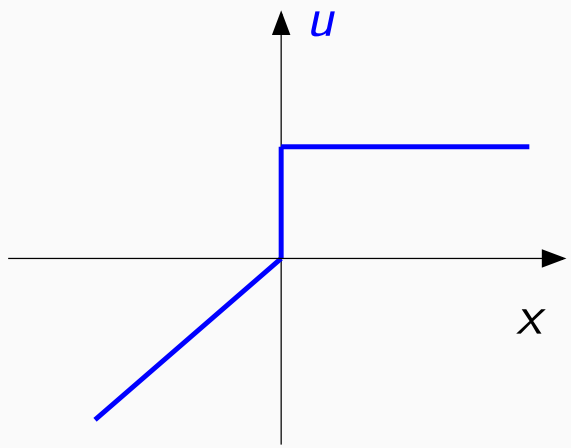
Subdifferential of function: definition

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

$$\partial f : \mathcal{H} \to 2^{\mathcal{H}}$$

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



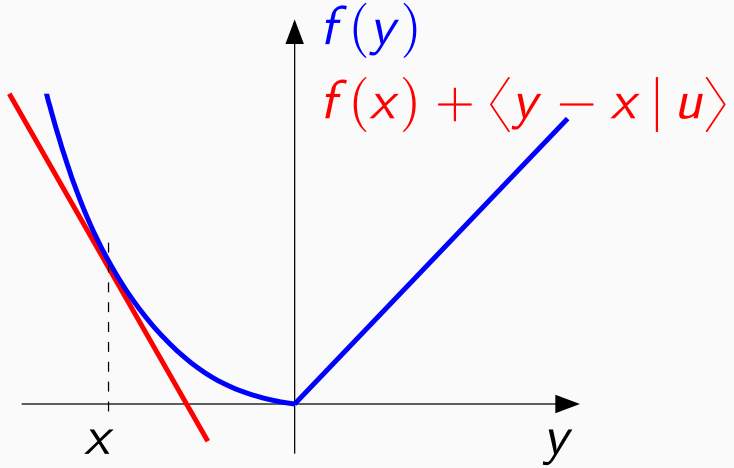


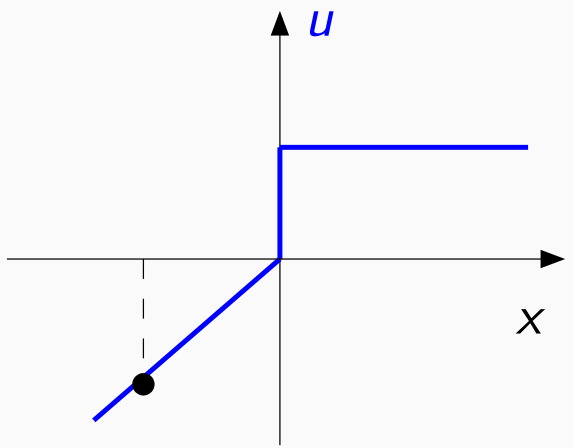
Subdifferential of function: definition

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

$$\partial f : \mathcal{H} \to 2^{\mathcal{H}}$$

$$x \to \{u \in \mathcal{H} \,|\, (\forall y \in \mathcal{H})\ \langle y - x | u \rangle + f(x) \leqslant f(y)\}$$



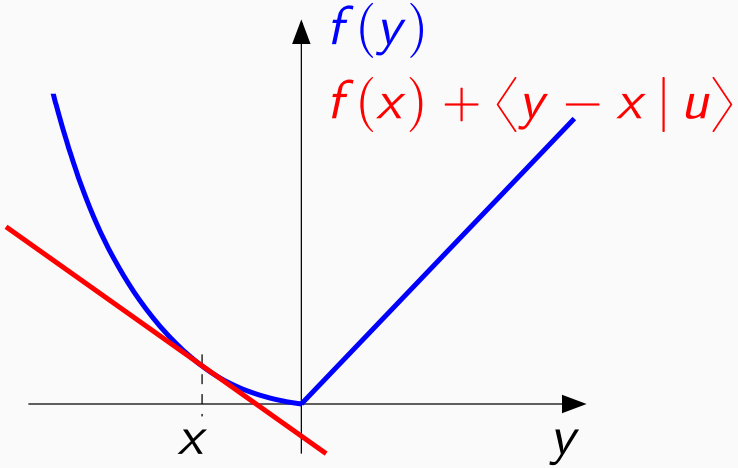


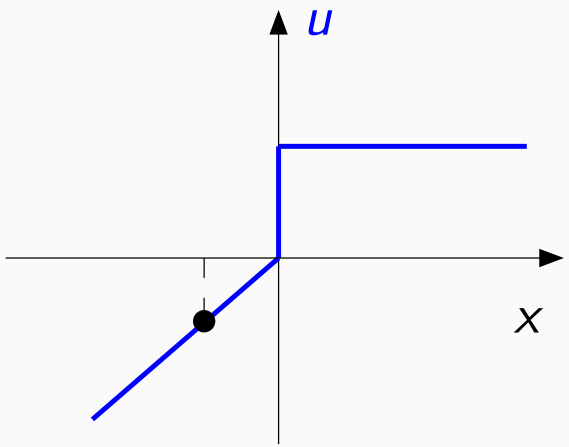
Subdifferential of function: definition

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

$$\partial f : \mathcal{H} \to 2^{\mathcal{H}}$$

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



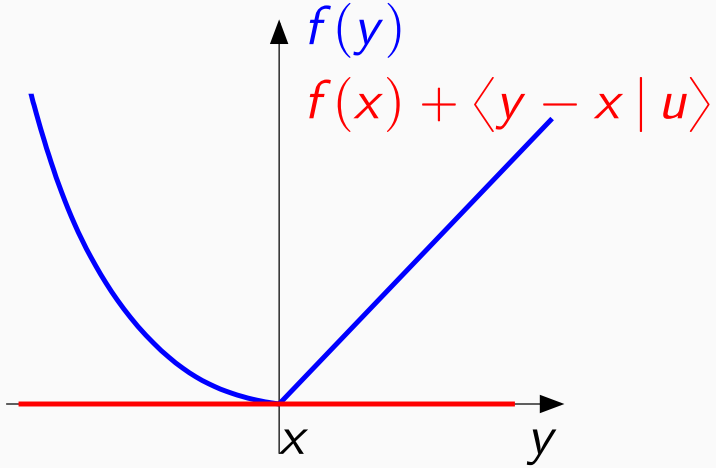


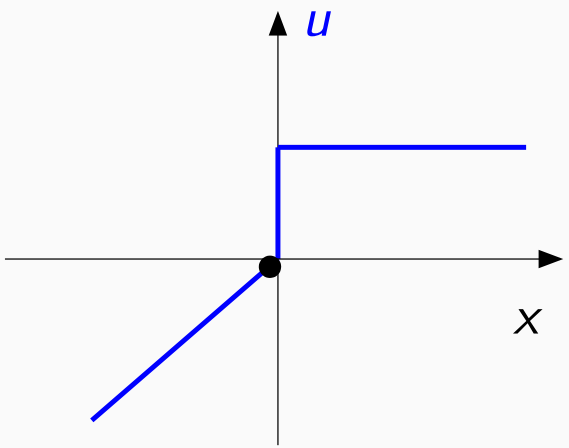
Subdifferential of function: definition

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

$$\partial f : \mathcal{H} \to 2^{\mathcal{H}}$$

$$x \to \{u \in \mathcal{H} \,|\, (\forall y \in \mathcal{H})\ \langle y - x | u \rangle + f(x) \leqslant f(y)\}$$



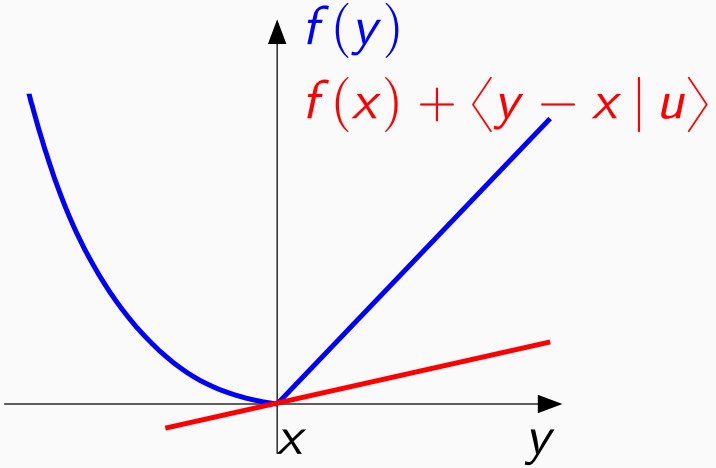


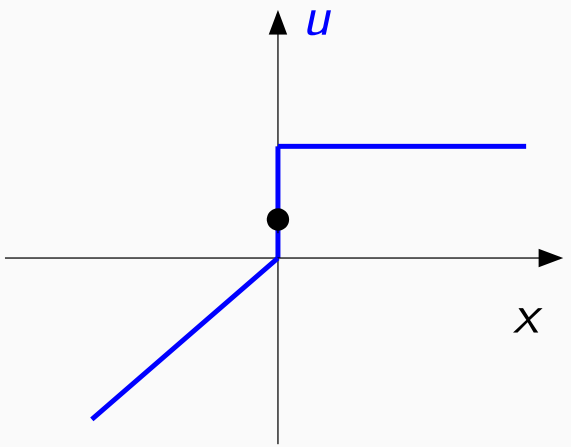
Subdifferential of function: definition

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

$$\partial f : \mathcal{H} \to 2^{\mathcal{H}}$$

$$x \to \{u \in \mathcal{H} \,|\, (\forall y \in \mathcal{H})\ \langle y - x | u \rangle + f(x) \leqslant f(y)\}$$



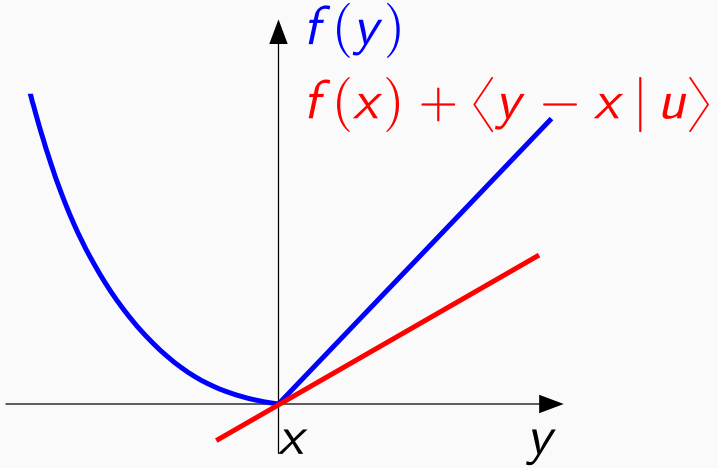


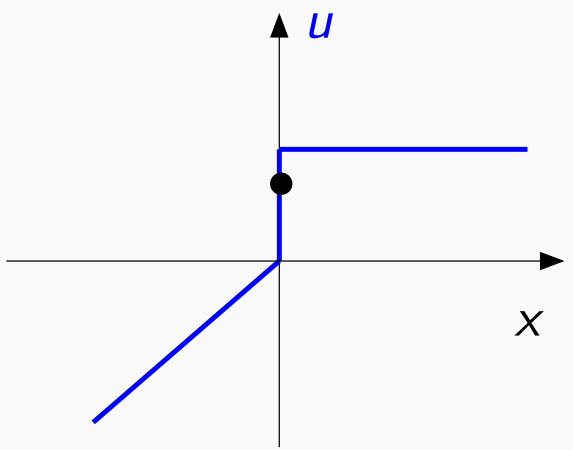
Subdifferential of function: definition

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

$$\partial f : \mathcal{H} \to 2^{\mathcal{H}}$$

$$x \to \{u \in \mathcal{H} \,|\, (\forall y \in \mathcal{H})\ \langle y - x | u \rangle + f(x) \leqslant f(y)\}$$



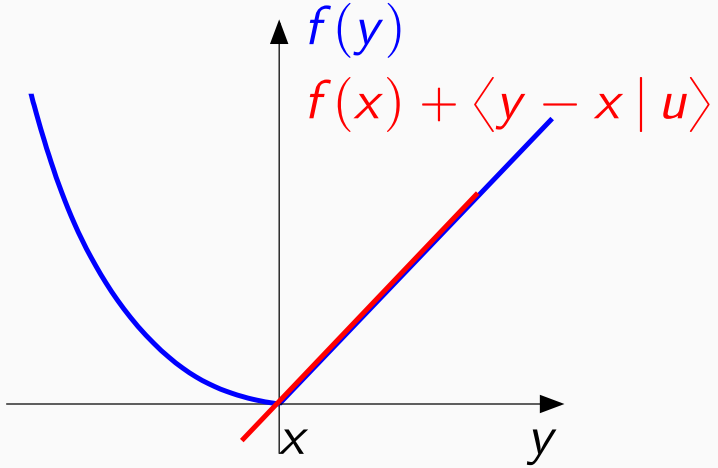


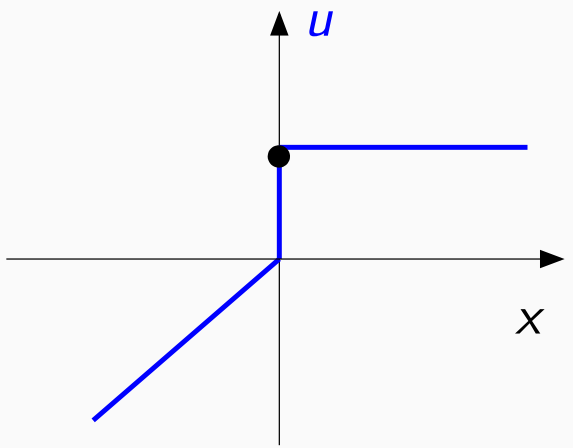
Subdifferential of function: definition

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

$$\partial f : \mathcal{H} \to 2^{\mathcal{H}}$$

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



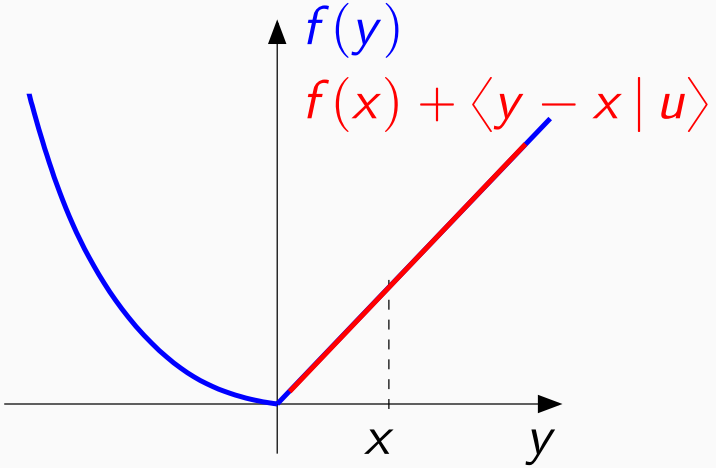


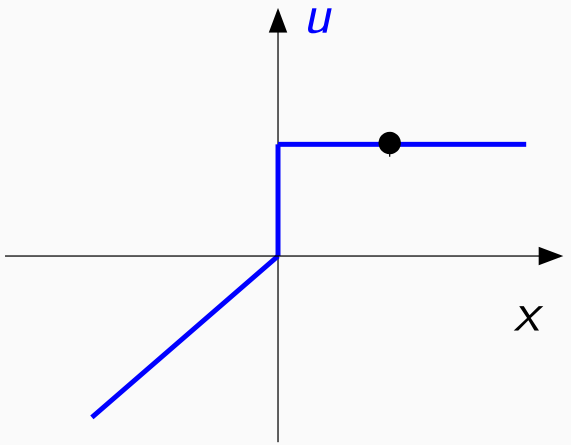
Subdifferential of function: definition

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

$$\partial f : \mathcal{H} \to 2^{\mathcal{H}}$$

$$x \to \{u \in \mathcal{H} \,|\, (\forall y \in \mathcal{H})\ \langle y - x | u \rangle + f(x) \leqslant f(y)\}$$

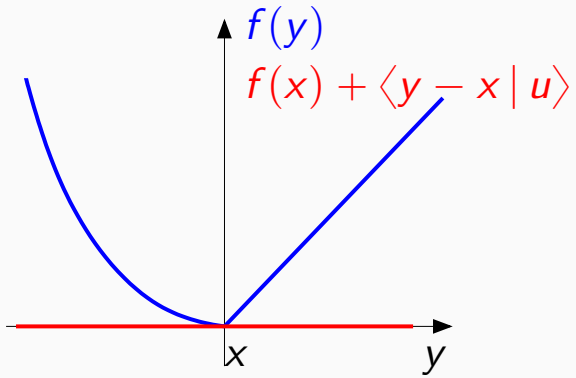


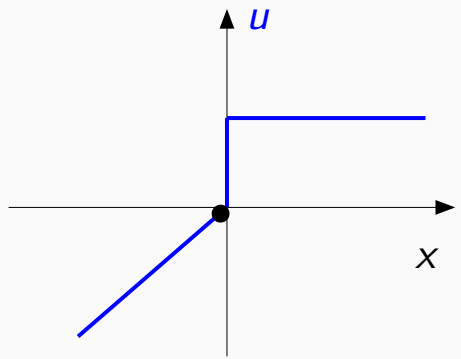


Subdifferential of a function: properties

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

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





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

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

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

Subdifferential of a convex function: properties

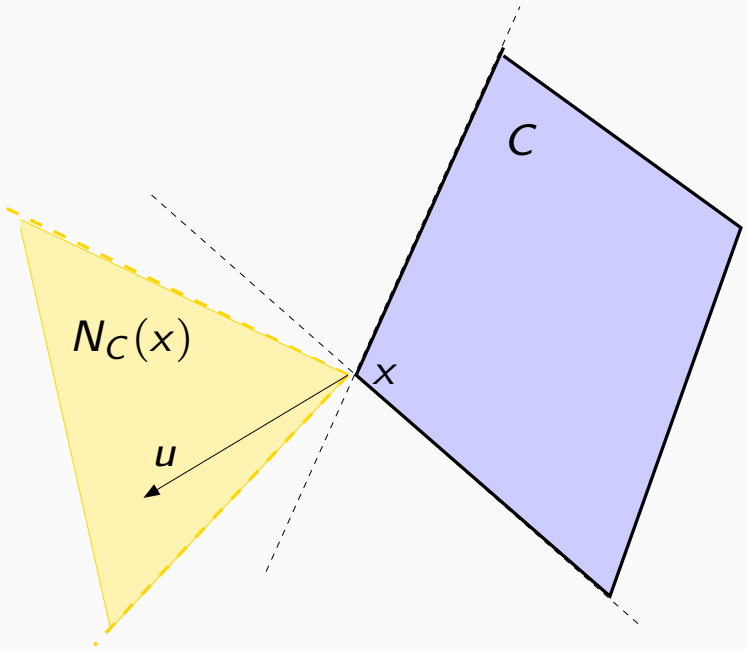
If $f : \mathcal{H} \to]-\infty, +\infty]$ is convex and it is Gâteaux differentiable at x, then

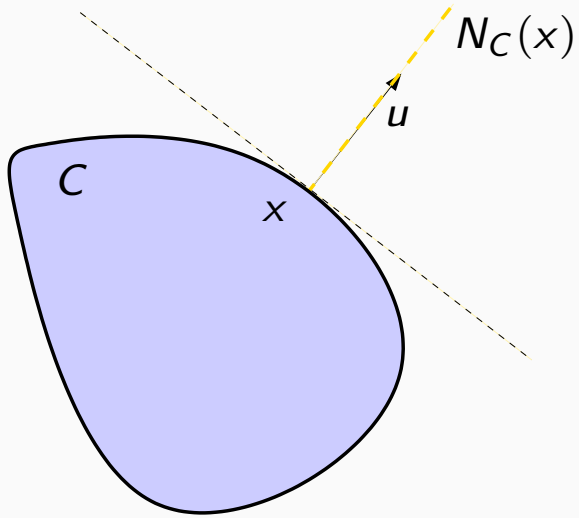
$$\partial f(x) = \{\nabla f(x)\}$$

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

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

Then,

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

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

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

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 $f_i : \mathcal{H}_i \to]-\infty, +\infty]$ be a proper function. Let

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

Then,

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

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 $(\forall i \in I\backslash\{j\})\ y_i = x_i \in dom\, f_i$, 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

$\rightarrow$ ℓ_1-**norm**

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

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

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

Mathematical foundations in deep learning
Part IV: Optimization – Conjugate

Nelly Pustelnik

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

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

Conjugate

Adrien-Marie Legendre
(1752–1833)

Werner Fenchel
(1905–1988)

Conjugate

Adrien-Marie Legendre
(1752–1833)

Werner Fenchel
(1905–1988)

Conjugate: definition

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

The conjugate of f is the function $f^* : \mathcal{H} \to [-\infty, +\infty]$ such that

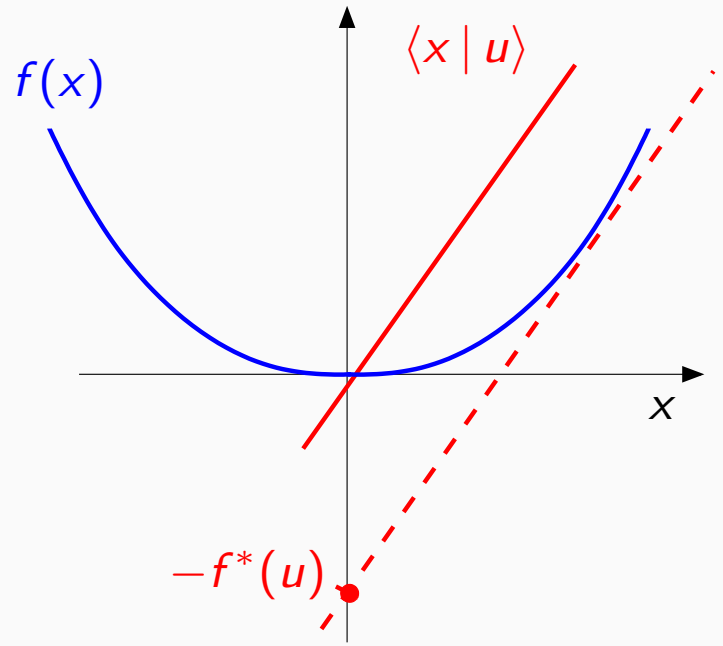
$$(\forall u \in \mathcal{H}) \qquad f^*(u) = \sup_{x \in \mathcal{H}} \big(\langle x \mid u \rangle - f(x) \big) .$$

Conjugate: definition

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

The conjugate of f is the function $f^* : \mathcal{H} \to [-\infty, +\infty]$ such that

$$(\forall u \in \mathcal{H}) \qquad f^*(u) = \sup_{x \in \operatorname{dom} f} \big(\langle x \mid u \rangle - f(x) \big)$$

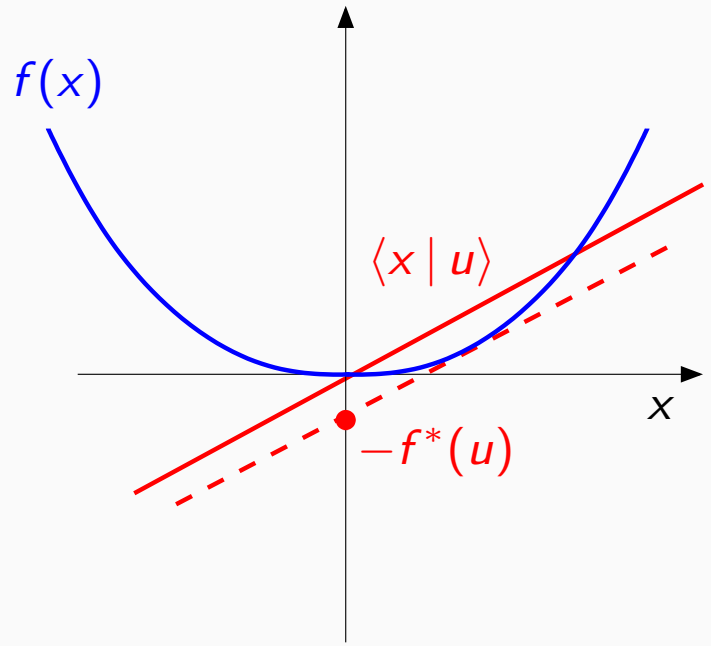


Conjugate: definition

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

The conjugate of f is the function $f^* : \mathcal{H} \to [-\infty, +\infty]$ such that

$$(\forall u \in \mathcal{H}) \qquad f^*(u) = \sup_{x \in \operatorname{dom} f} \big(\langle x \mid u \rangle - f(x) \big)$$

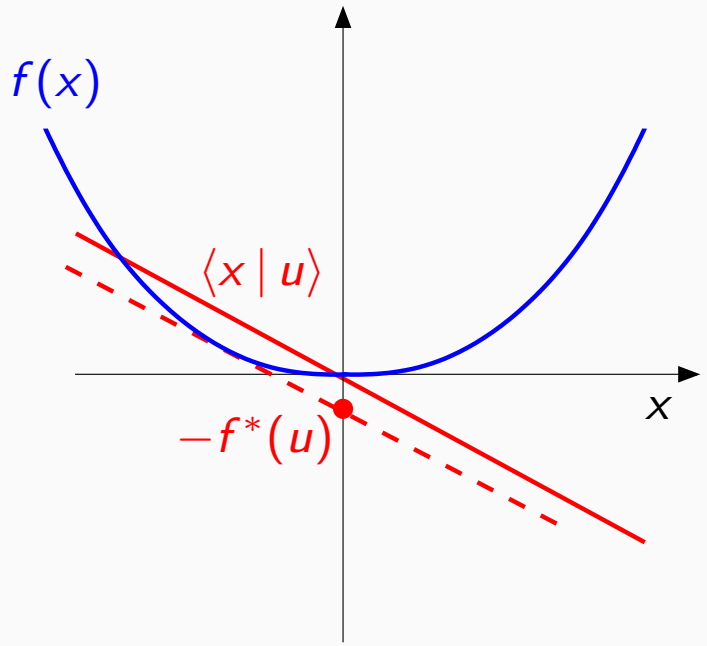


Conjugate: definition

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

The conjugate of f is the function $f^* : \mathcal{H} \to [-\infty, +\infty]$ such that

$$(\forall u \in \mathcal{H}) \qquad f^*(u) = \sup_{x \in \operatorname{dom} f} \big(\langle x \mid u \rangle - f(x) \big)$$



Conjugate: definition

Let $\mathcal{H}$ be a Hilbert space and $f : \mathcal{H} \to]-\infty, +\infty]$.
The conjugate of f is the function $f^* : \mathcal{H} \to [-\infty, +\infty]$ such that

$$(\forall u \in \mathcal{H}) \qquad f^*(u) = \sup_{x \in \operatorname{dom} f} \big(\langle x \mid u \rangle - f(x) \big)$$

Examples :

- $f = \frac{1}{2}\|\cdot\|^2 \Rightarrow f^* = \frac{1}{2}\|\cdot\|^2$

  Proof : For every $(x, u) \in \mathcal{H}^2$, $\langle x \mid u \rangle - \frac{1}{2}\|x\|^2 = \frac{1}{2}\|u\|^2 - \frac{1}{2}\|u - x\|^2$ is maximum at $x = u$.
  Consequently, $f^*(u) = \frac{1}{2}\|u\|^2$.

Conjugate: definition

Let $\mathcal{H}$ be a Hilbert space and $f : \mathcal{H} \to]-\infty, +\infty]$.
The conjugate of f is the function $f^* : \mathcal{H} \to [-\infty, +\infty]$ such that

$$(\forall u \in \mathcal{H}) \qquad f^*(u) = \sup_{x \in \operatorname{dom} f} \big(\langle x \mid u \rangle - f(x) \big)$$

Examples :

- $f = \frac{1}{2}\|\cdot\|^2 \Rightarrow f^* = \frac{1}{2}\|\cdot\|^2$.

- $(\forall x \in \mathbb{R}^N)\ f(x) = \frac{1}{q}\|x\|_q^q$ with $q \in]1, +\infty[$
  $\Rightarrow (\forall u \in \mathbb{R}^N)\ f^*(u) = \frac{1}{q^*}\|u\|_{q^*}^{q^*}$ with $\frac{1}{q} + \frac{1}{q^*} = 1$

Conjugate: definition

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

The conjugate of f is the function $f^* : \mathcal{H} \to [-\infty, +\infty]$ such that

$$(\forall u \in \mathcal{H}) \qquad f^*(u) = \sup_{x \in \operatorname{dom} f} \big(\langle x \mid u \rangle - f(x) \big)$$

- If f is even, then f^* is even.

Conjugate: definition

Let $\mathcal{H}$ be a Hilbert space and $f : \mathcal{H} \to]-\infty, +\infty]$.
The conjugate of f is the function $f^* : \mathcal{H} \to [-\infty, +\infty]$ such that

$$(\forall u \in \mathcal{H}) \qquad f^*(u) = \sup_{x \in \operatorname{dom} f} \big(\langle x \mid u \rangle - f(x) \big)$$

- If f is even, then f^* is even.
- For every $\alpha \in]0, +\infty[$, $(\alpha f)^* = \alpha f^*(\cdot/\alpha)$.
- For every $(y, v) \in \mathcal{H}^2$ et $\alpha \in \mathbb{R}$,
  $\big(f(\cdot - y) + \langle \cdot \mid v \rangle + \alpha\big)^* = f^*(\cdot - v) + \langle y \mid \cdot - v \rangle - \alpha$.
- Let $\mathcal{G}$ be a Hilbert space and $L \in \mathcal{B}(\mathcal{G}, \mathcal{H})$ be an isomorphism.
  $(f \circ L)^* = f^* \circ (L^{-1})^*$.
- f^* is l.s.c. and convex.

Conjugate: definition

Let $\mathcal{H}$ be a Hilbert space and $f : \mathcal{H} \to]-\infty, +\infty]$.
The conjugate of f is the function $f^* : \mathcal{H} \to [-\infty, +\infty]$ such that

$$(\forall u \in \mathcal{H}) \qquad f^*(u) = \sup_{x \in \operatorname{dom} f} \big(\langle x \mid u \rangle - f(x) \big)$$

Moreau-Fenchel theorem

Let $\mathcal{H}$ be a Hilbert space and $f : \mathcal{H} \to]-\infty, +\infty]$ be a proper function.

$$f \text{ is l.s.c. and convex} \Leftrightarrow f^{**} = f.$$

Conjugate: definition

Let $\mathcal{H}$ be a Hilbert space and $f : \mathcal{H} \to]-\infty, +\infty]$.
The conjugate of f is the function $f^* : \mathcal{H} \to [-\infty, +\infty]$ such that

$$(\forall u \in \mathcal{H}) \qquad f^*(u) = \sup_{x \in \operatorname{dom} f} \big(\langle x \mid u \rangle - f(x) \big)$$

Moreau-Fenchel theorem

Let $\mathcal{H}$ be a Hilbert space and $f : \mathcal{H} \to]-\infty, +\infty]$ be a proper function.

$$f \text{ is l.s.c. and convex} \Leftrightarrow f^{**} = f.$$

- Consequence: If $f \in \Gamma_0(\mathbb{R})$, then f^* is proper, hence $f^* \in \Gamma_0(\mathbb{R})$.

Conjugate: properties

Fenchel-Young inequality: If f is proper, then

1. $\big(\forall (x,u) \in \mathcal{H}^2\big) \qquad f(x) + f^*(u) \geq \langle x \mid u \rangle$

2. $\big(\forall (x,u) \in \mathcal{H}^2\big) \qquad u \in \partial f(x) \Leftrightarrow \quad f(x) + f^*(u) = \langle x \mid u \rangle.$

If $f \in \Gamma_0(\mathcal{H})$, then

$$\big(\forall (x,u) \in \mathcal{H}^2\big) \qquad u \in \partial f(x) \;\Leftrightarrow\; x \in \partial f^*(u)\,.$$

Conjugate: properties

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 $f_i : \mathcal{H}_i \to]-\infty, +\infty]$. Let

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

Then,

$$(\forall u = (u_i)_{i\in I} \in \mathcal{H}) \qquad f^*(u) = \sum_{i\in I} f_i^*(u_i)\,.$$

Conjugate: properties

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 $f_i : \mathcal{H}_i \to]-\infty, +\infty]$. Let

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

Then,

$$\big(\forall u = (u_i)_{i\in I} \in \mathcal{H}\big) \qquad \boxed{f^*(u) = \sum_{i\in I} f_i^*(u_i)}\,.$$

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

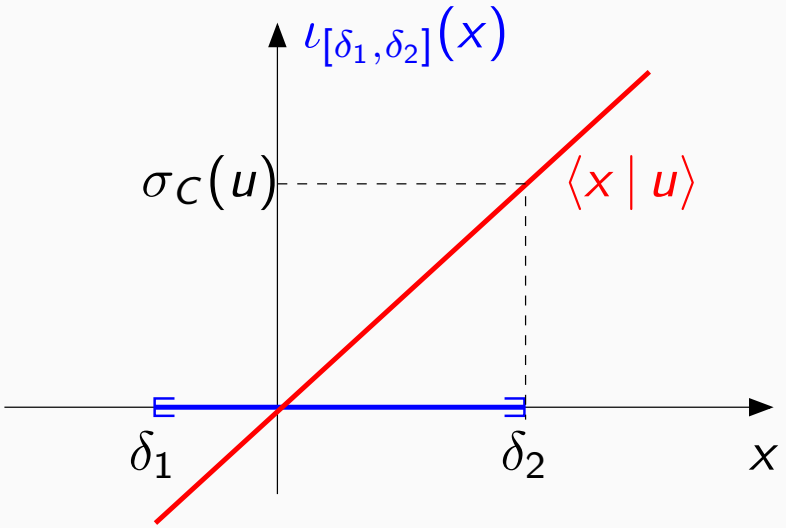
$$\begin{aligned} f^*(u) &= \sup_{x\in\mathcal{H}} \langle x \mid u\rangle - f(x) \\ &= \sup_{x=(x_i)_{i\in I}\in\mathcal{H}} \sum_{i\in I} \langle x_i \mid u_i\rangle - f_i(x_i) \\ &= \sum_{i\in I} \sup_{x_i\in\mathcal{H}_i} \langle x_i \mid u_i\rangle - f_i(x_i) \\ &= \sum_{i\in I} f_i^*(u_i). \end{aligned}$$

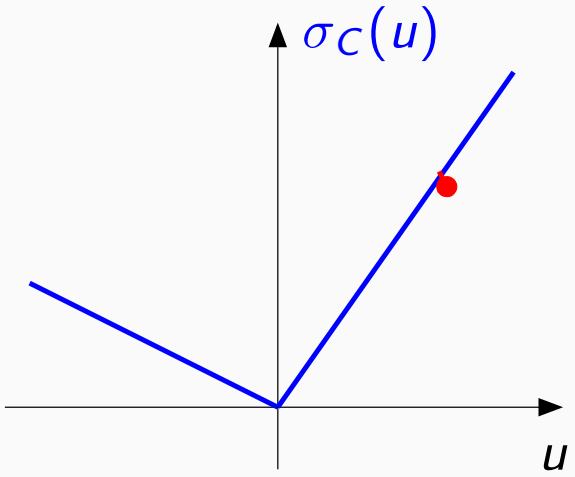
Conjugate: example

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

σ_C is the support function of C if

$$(\forall u \in \mathcal{H}) \qquad \begin{aligned} \sigma_C(u) &= \sup_{x \in C} \langle x \mid u \rangle \\ &= \iota_C^*(u). \end{aligned}$$



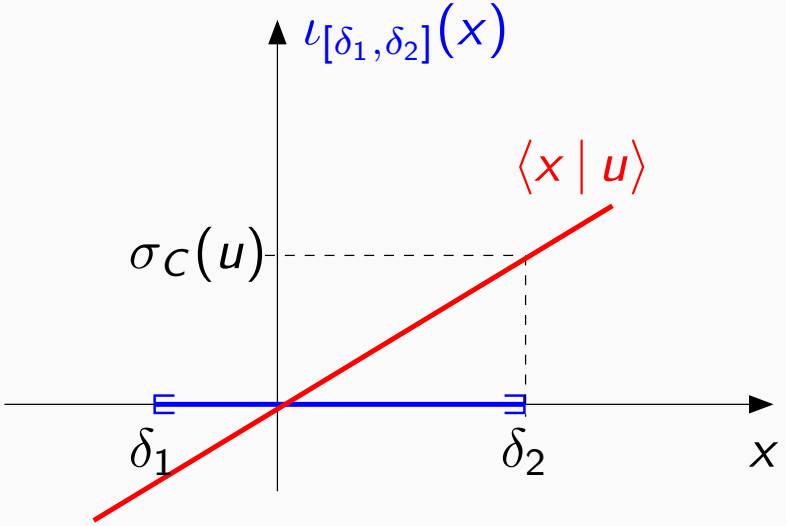


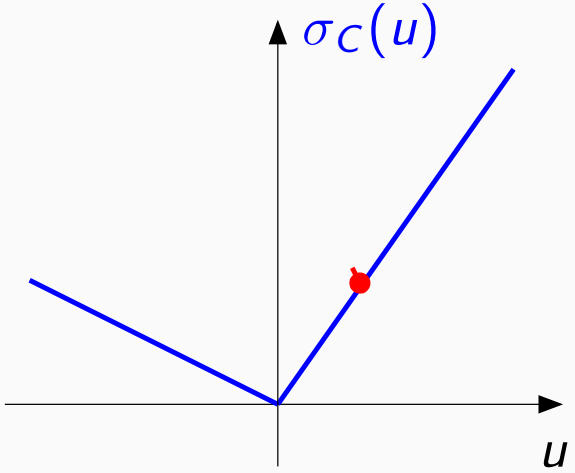
Conjugate: example

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

σ_C is the support function of C if

$$(\forall u \in \mathcal{H}) \qquad \begin{aligned} \sigma_C(u) &= \sup_{x \in C} \langle x \mid u \rangle \\ &= \iota_C^*(u). \end{aligned}$$



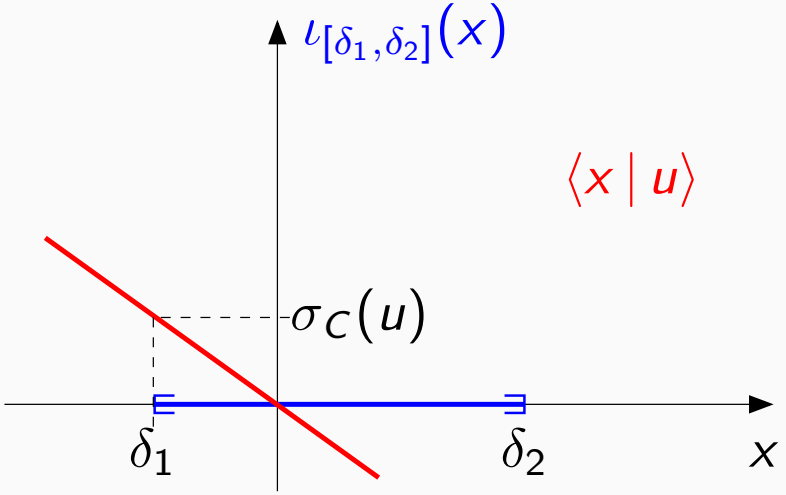


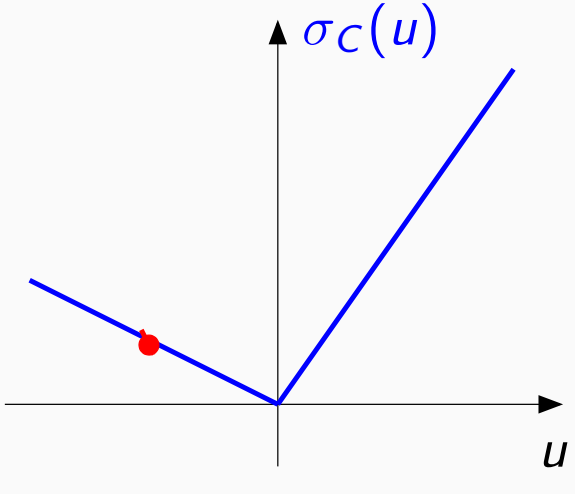
Conjugate: example

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

σ_C is the support function of C if

$$(\forall u \in \mathcal{H}) \qquad \begin{aligned} \sigma_C(u) &= \sup_{x \in C} \langle x \mid u \rangle \\ &= \iota_C^*(u). \end{aligned}$$





Proximity operator: support function

Support function :

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

$$(\forall x \in \mathcal{H}) \qquad \operatorname{prox}_{\sigma_C} = \mathrm{Id} - P_C.$$

Conjugate: example

- Let $f : \mathbb{R} \to]-\infty, +\infty] : x \mapsto \begin{cases} \delta_1 x & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ \delta_2 x & \text{if } x > 0 \end{cases}$
  with $-\infty \leq \delta_1 < \delta_2 \leq +\infty$.
  Then, $f = \sigma_C$ where C is the closed real interval such that $\inf C = \delta_1$ et $\sup C = \delta_2$.

Conjugate: example

- Let $f : \mathbb{R} \to]-\infty, +\infty] : x \mapsto \begin{cases} \delta_1 x & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ \delta_2 x & \text{if } x > 0 \end{cases}$
  with $-\infty \leq \delta_1 < \delta_2 \leq +\infty$.
  Then, $f = \sigma_C$ where C is the closed real interval such that $\inf C = \delta_1$ et $\sup C = \delta_2$.

- Let f be a ℓ^q norm of $\mathbb{R}^N$ with $q \in [1, +\infty]$.
  We have $f = \sigma_C$ where

$$C = \left\{ y \in \mathbb{R}^N \;\middle|\; \|y\|_{q^*} \leq 1 \right\} \qquad \text{with } \frac{1}{q} + \frac{1}{q^*} = 1.$$

Conjugate: example

- Let $f : \mathbb{R} \to]-\infty, +\infty] : x \mapsto \begin{cases} \delta_1 x & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ \delta_2 x & \text{if } x > 0 \end{cases}$
  with $-\infty \leq \delta_1 < \delta_2 \leq +\infty$.
  Then, $f = \sigma_C$ where C is the closed real interval such that $\inf C = \delta_1$ et $\sup C = \delta_2$.

- Let f be a ℓ^q norm of $\mathbb{R}^N$ with $q \in [1, +\infty]$.
  We have $f = \sigma_C$ where

$$C = \left\{ y \in \mathbb{R}^N \;\middle|\; \|y\|_{q^*} \leq 1 \right\} \qquad \text{with } \frac{1}{q} + \frac{1}{q^*} = 1.$$

  Particular case: ℓ^1 norm of $\mathbb{R}^N$: $C = [-1, 1]^N$.

Mathematical foundations in deep learning
Part V: Optimization – 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 $\quad \widehat{x} \in \underset{x \in \mathbb{R}^N}{\operatorname{Argmin}}\, f(x)$.

- 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

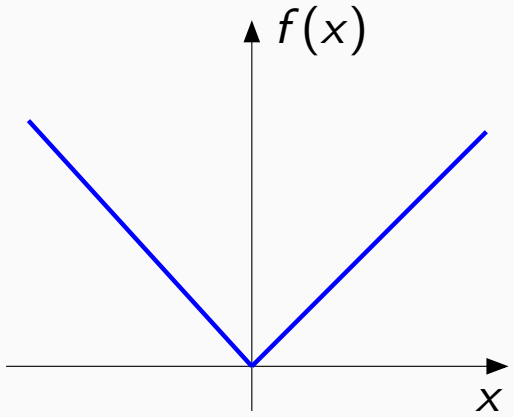
$$(\forall n \in \mathbb{N}) \qquad x_{n+1} = x_n - \gamma_n u_n \quad \text{with} \quad u_n \in \partial f(x_{n+1})$$

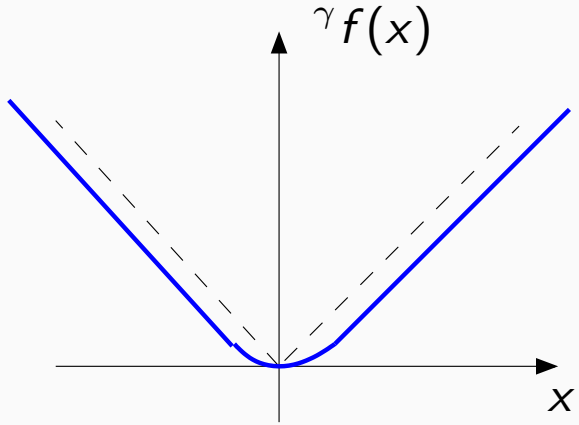
$\rightarrow$ Convergence insured when $\sum_{n=0}^{+\infty} \gamma_n = +\infty \Rightarrow$ **Proximity operator**.

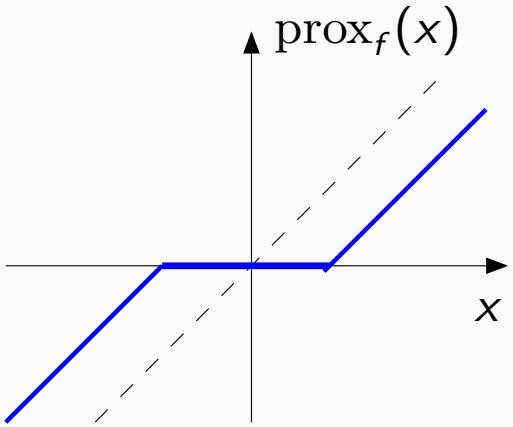
Proximity operator: definition

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

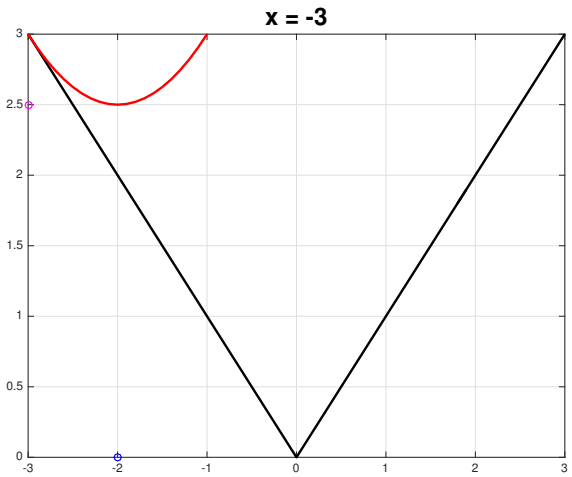
- The **Moreau envelope** of f of parameter $\gamma \in]0, +\infty[$ is
$$^{\gamma}f : \mathcal{H} \to \mathbb{R} : x \mapsto \inf_{y \in \mathcal{H}} f(y) + \frac{1}{2\gamma}\|y - x\|^2.$$
- The **proximity operator** of f is
$$\operatorname{prox}_f : \mathcal{H} \to \mathcal{H} : x \mapsto \operatorname*{argmin}_{y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





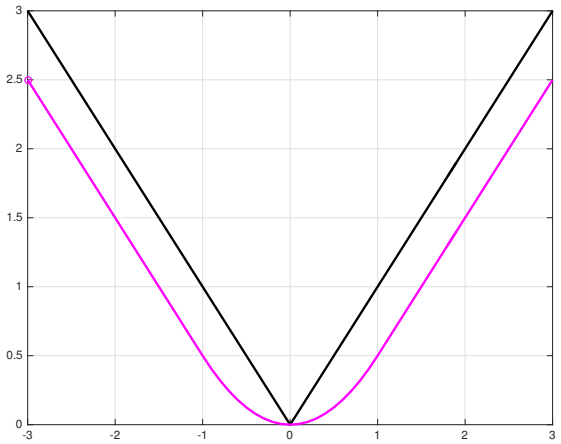


Proximity operator: definition



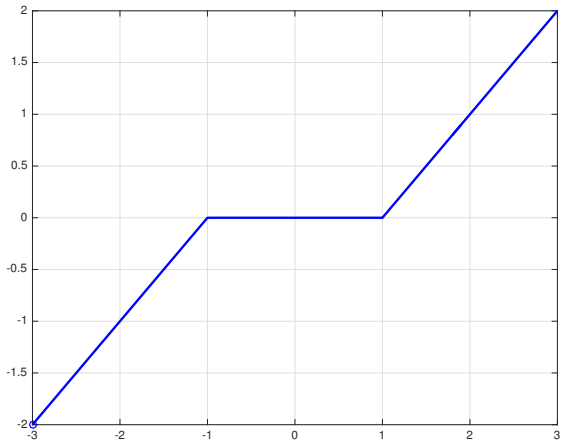
$f(y) = |y|$

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



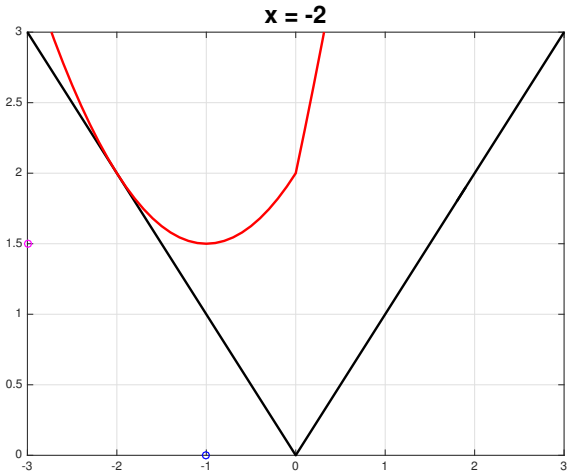
$f(x) = |x|$

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



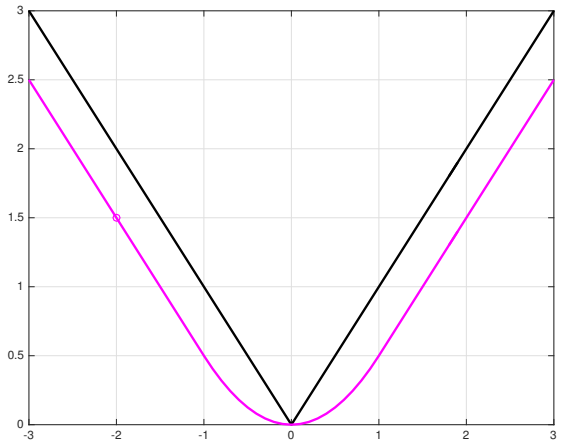
$\mathrm{prox}_f(x) = \underset{y \in \mathcal{H}}{\mathrm{argmin}} \; g(y;x)$

Proximity operator: definition



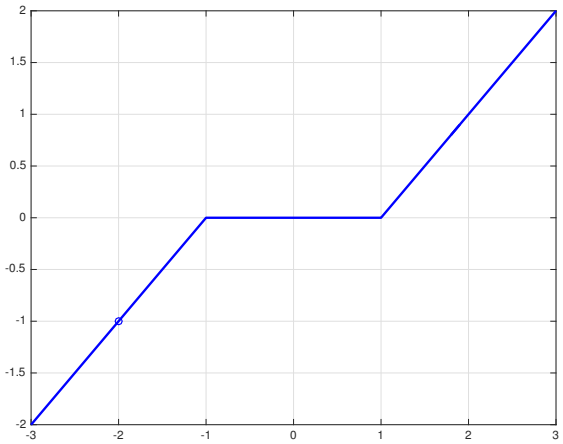
$f(y) = |y|$

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



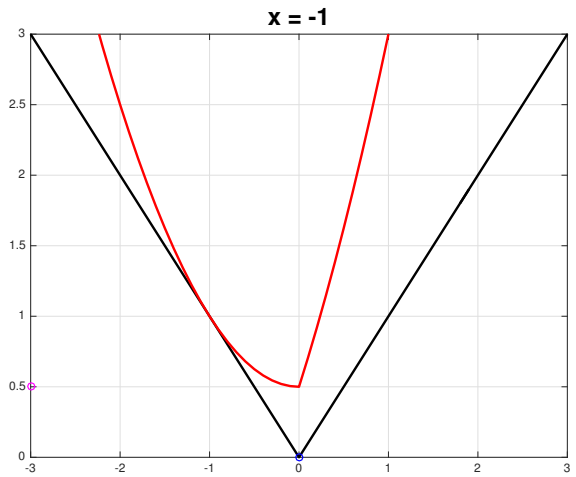
$f(x) = |x|$

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



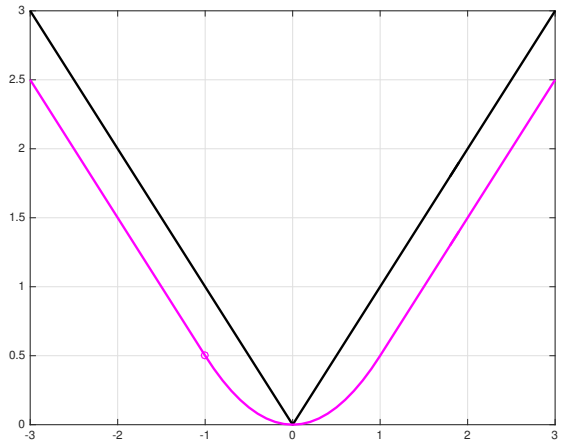
$\mathrm{prox}_f(x) = \underset{y \in \mathcal{H}}{\mathrm{argmin}} \; g(y;x)$

Proximity operator: definition



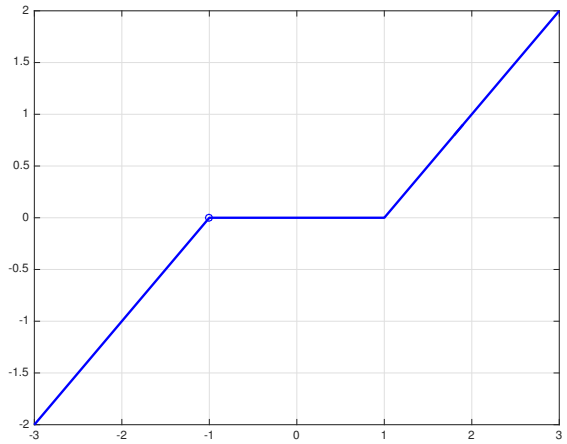
$f(y) = |y|$

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



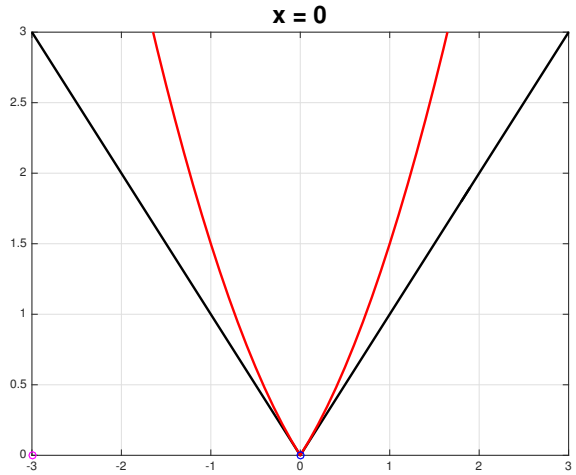
$f(x) = |x|$

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



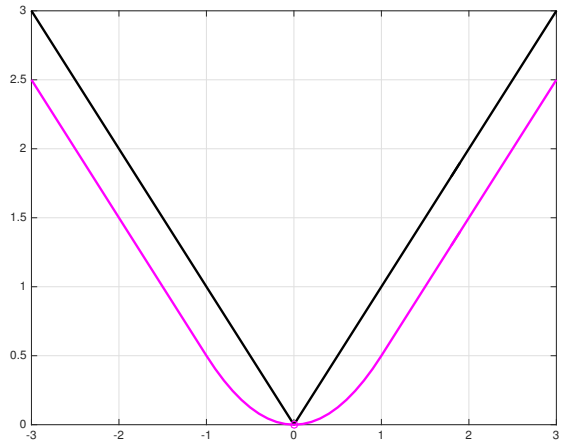
$\mathrm{prox}_f(x) = \underset{y \in \mathcal{H}}{\mathrm{argmin}} \; g(y;x)$

Proximity operator: definition



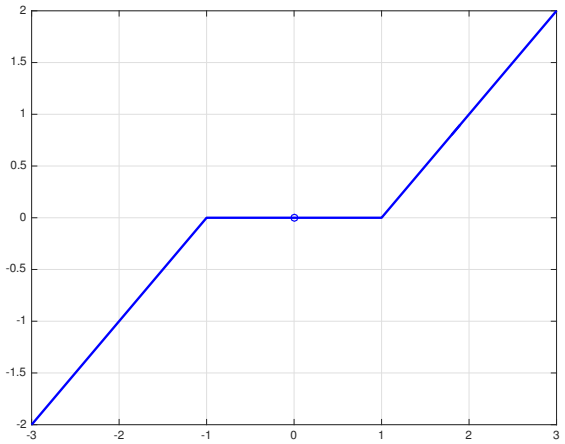
$f(y) = |y|$

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



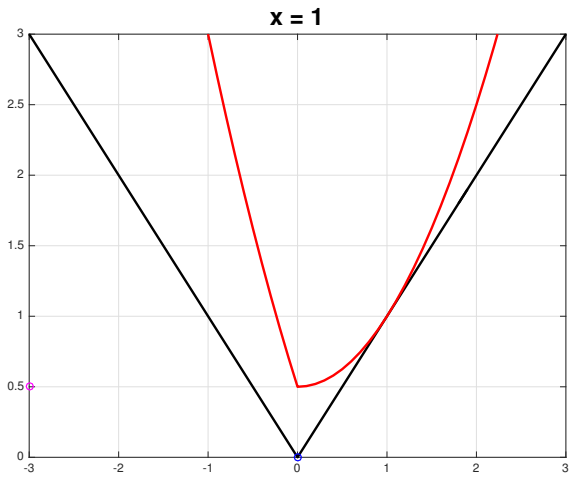
$f(x) = |x|$

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



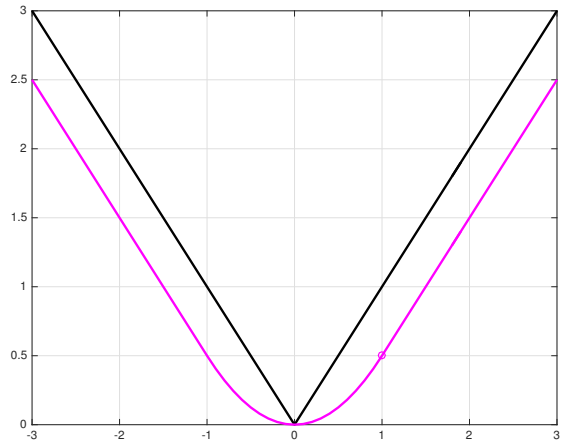
$\mathrm{prox}_f(x) = \underset{y \in \mathcal{H}}{\mathrm{argmin}} \; g(y;x)$

Proximity operator: definition



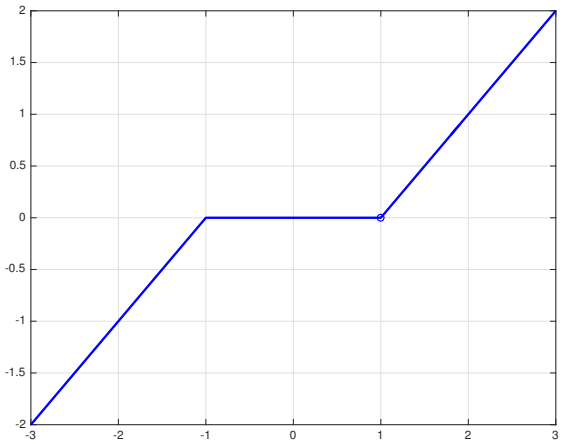
$f(y) = |y|$

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



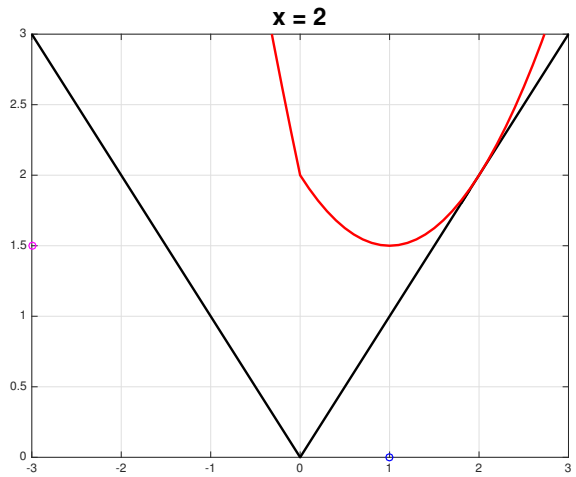
$f(x) = |x|$

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



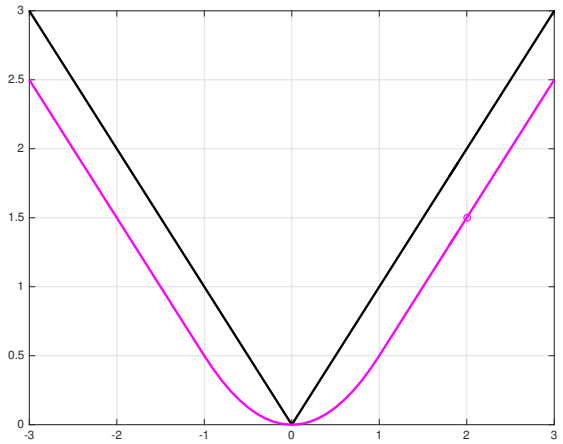
$\mathrm{prox}_f(x) = \underset{y \in \mathcal{H}}{\mathrm{argmin}} \; g(y;x)$

Proximity operator: definition



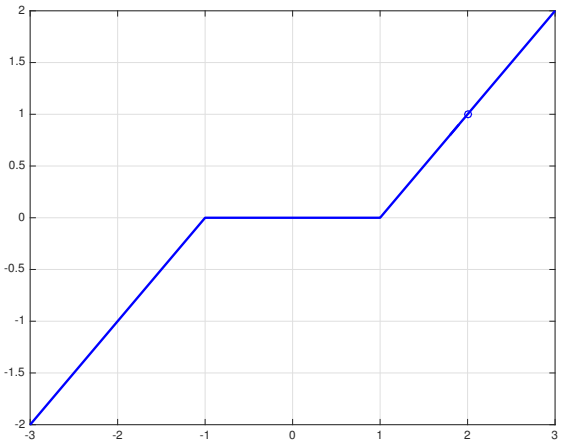
$f(y) = |y|$

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



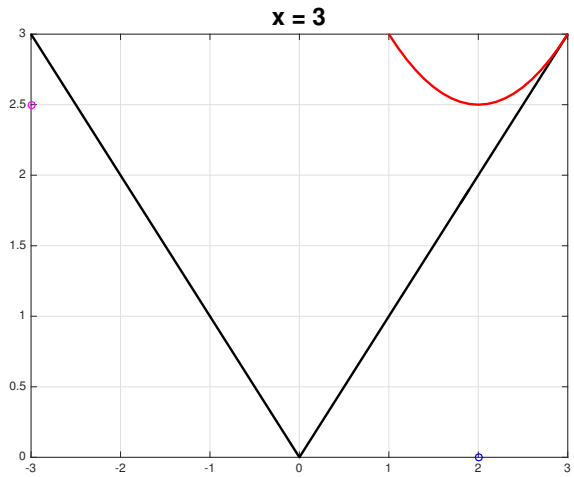
$f(x) = |x|$

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



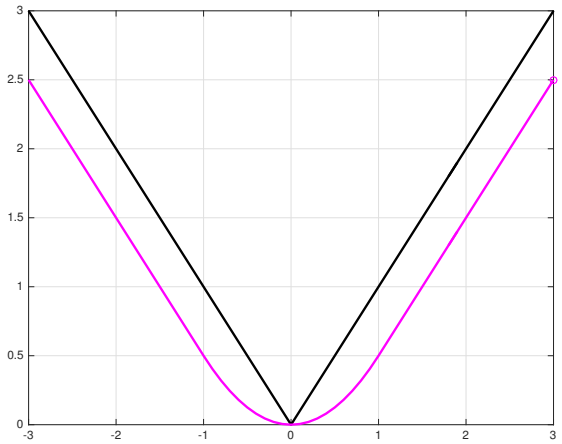
$\mathrm{prox}_f(x) = \underset{y \in \mathcal{H}}{\mathrm{argmin}} \; g(y;x)$

Proximity operator: definition



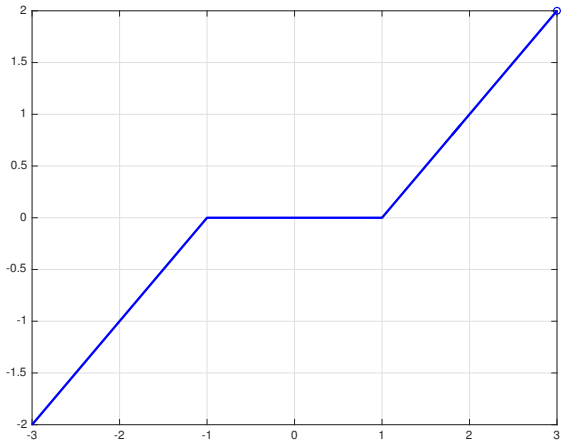
$f(y) = |y|$

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



$f(x) = |x|$

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



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

Proximity operator: characterization

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

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

Proximity operator: characterization

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

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

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

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

Proximity operator: characterization

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

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

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

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

Proximity operator: existence and uniqueness

Let $f \in \Gamma_0(\mathcal{H})$ and $\gamma \in \,]0, +\infty[$.
For every $x \in \mathcal{H}$, there exists a unique vector $p \in \mathcal{H}$ such that

$$f(p) + \frac{1}{2\gamma}\|p - x\|^2 = \inf_{y \in \mathcal{H}} f(y) + \frac{1}{2\gamma}\|y - x\|^2 .$$

Proof: $f \in \Gamma_0(\mathcal{H}) \Leftrightarrow f^* \in \Gamma_0(\mathcal{H})$. Thus, there exists $u \in \mathcal{H}$ such that $f^*(u) \in \mathbb{R}$. According to Fenchel-Young inequality, we have

$$(\forall y \in \mathcal{H}) \qquad f(y) \geq \langle u \mid y \rangle - f^*(u).$$

Then, $f(y) + (2\gamma)^{-1}\|y - x\|^2 \to +\infty$ when $\|y\| \to +\infty$.
Furthermore $(2\gamma)^{-1}\|\cdot - x\|^2$ being strictly convex, $f + (2\gamma)^{-1}\|\cdot - x\|^2$ is a strictly convex coercive function.

Proximity operator: examples

Projection :

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

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

Proximity operator: examples

Power q function with $q \geq 1$:

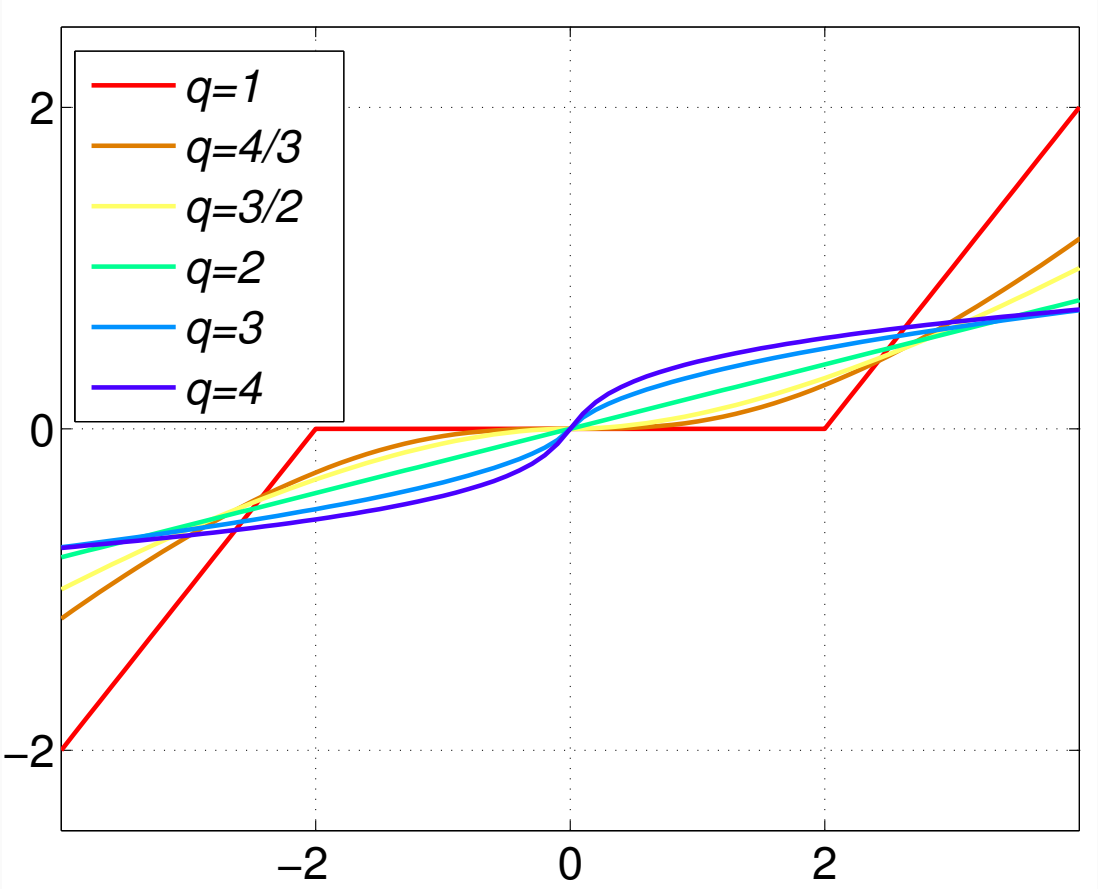
Let $\chi > 0$, $q \in [1, +\infty[$ and $\varphi \colon \mathbb{R} \to]-\infty, +\infty]$: $\eta \mapsto \chi|\xi|^q$.

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 $z \in \mathcal{G}$.

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

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

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

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

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

for some $\tau > 0$, where

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

Proximity operator: properties

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

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

Proximity operator: properties

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

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

then

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

Proximity operator: support function

Support function :

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

$$(\forall x \in \mathcal{H}) \qquad \operatorname{prox}_{\sigma_C} = \mathrm{Id} - P_C.$$

Proximity operator: Moreau decomposition

Moreau decomposition formula

Let $\mathcal{H}$ be a Hilbert space, $f \in \Gamma_0(\mathcal{H})$ and $\gamma \in \,]0, +\infty[$.

$$(\forall x \in \mathcal{H}) \qquad \operatorname{prox}_{\gamma f^*} x = x - \gamma \operatorname{prox}_{\gamma^{-1} f}(\gamma^{-1} x)\,.$$

Proof:

$$\begin{aligned}
p = \operatorname{prox}_{\gamma f^*} x &\Leftrightarrow x - p \in \gamma \partial f^*(p) \\
&\Leftrightarrow p \in \partial f\left(\frac{x-p}{\gamma}\right) \\
&\Leftrightarrow \frac{x}{\gamma} - \frac{x-p}{\gamma} \in \frac{1}{\gamma}\partial f\left(\frac{x-p}{\gamma}\right) \\
&\Leftrightarrow \frac{x-p}{\gamma} = \operatorname{prox}_{\gamma^{-1} f}(\gamma^{-1} x) \\
&\Leftrightarrow p = x - \gamma \operatorname{prox}_{\gamma^{-1} f}(\gamma^{-1} x).
\end{aligned}$$

Proximity operator: Moreau decomposition

Moreau decomposition formula

Let $\mathcal{H}$ be a Hilbert space, $f \in \Gamma_0(\mathcal{H})$ and $\gamma \in \,]0, +\infty[$.

$$(\forall x \in \mathcal{H}) \qquad \operatorname{prox}_{\gamma f^*} x = x - \gamma \operatorname{prox}_{\gamma^{-1} f}(\gamma^{-1} x)\,.$$

Example: If $\mathcal{H} = \mathbb{R}^N$, $f = \frac{1}{q}\|\cdot\|_q^q$ with $q \in]1, +\infty[$, then $f^* = \frac{1}{q^*}\|\cdot\|_{q^*}^{q^*}$ with $1/q + 1/q^* = 1$, and

$$(\forall x \in \mathbb{R}^N) \qquad \operatorname{prox}_{\frac{\gamma}{q^*}\|\cdot\|_{q^*}^{q^*}} x = x - \gamma \operatorname{prox}_{\frac{1}{\gamma q}\|\cdot\|_q^q}(\gamma^{-1} x).$$

Proximity operator: properties

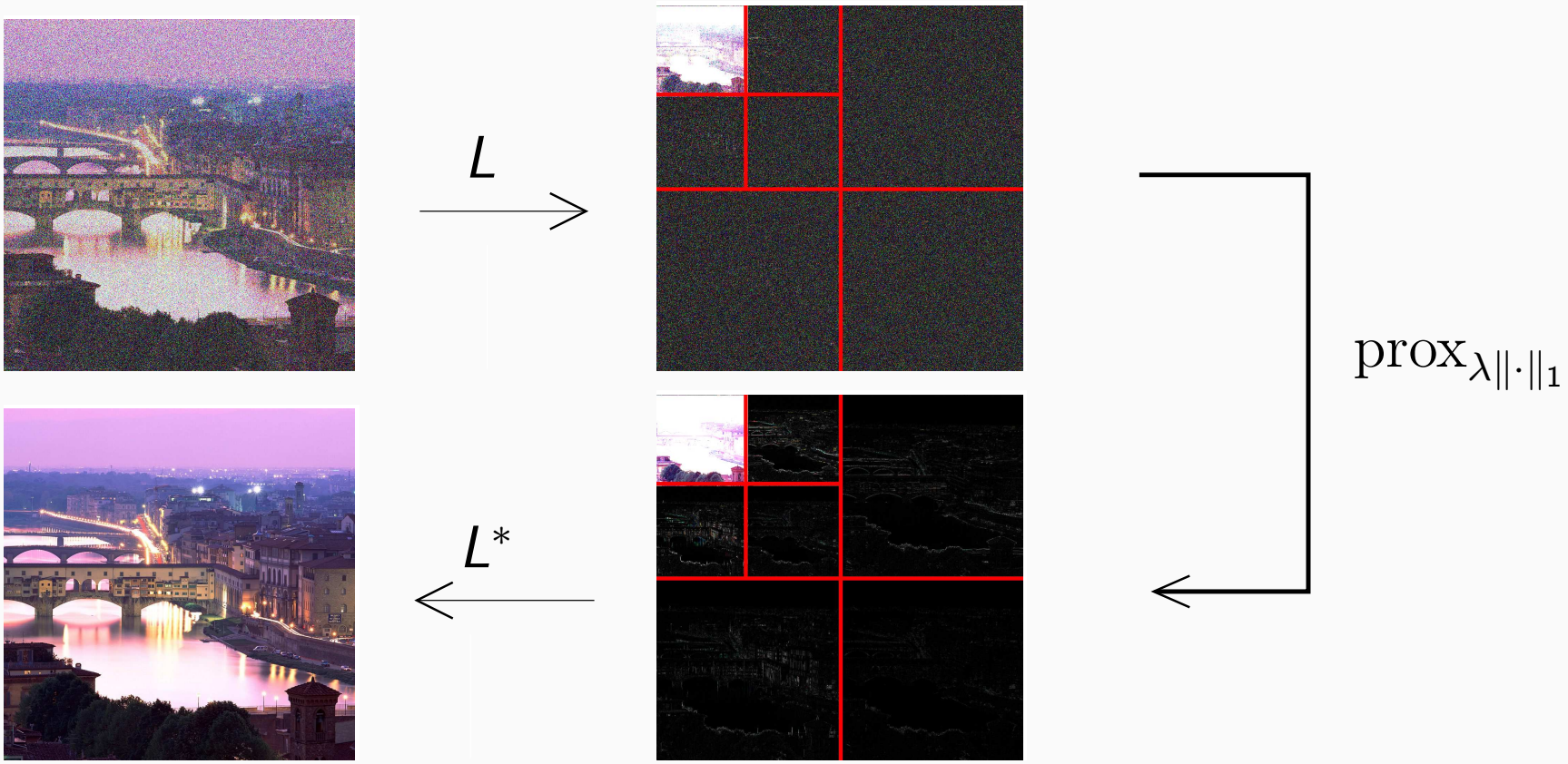
Let $\mathcal{H}$ and $\mathcal{G}$ be two Hilbert spaces. Let $f \in \Gamma_0(\mathcal{H})$ and $L \in \mathcal{B}(\mathcal{G}, \mathcal{H})$ such that $LL^* = \mu \mathrm{Id}$ where $\mu \in]0, +\infty[$. Then

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

Proximity operator: properties

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

- 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