Non-Smooth Convex Optimization in Data Sciences

Jalal Fadili

Normandie Université-ENSICAEN, GREYC

Mathematical coffees 2018

Normandie Université

cnrs

IUF

ENSI CAEN

ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

Outline

- Introduction.
- Non-smooth convex optimization.
  - Elements of convex analysis.
  - Elements of duality.
  - Optimality conditions.
- Proximal framework and operator splitting.
  - Proximal calculus.
  - Monotone operator splitting.
  - Sum of two functions.
  - Generalization to more than two functions.
- Take-away messages.

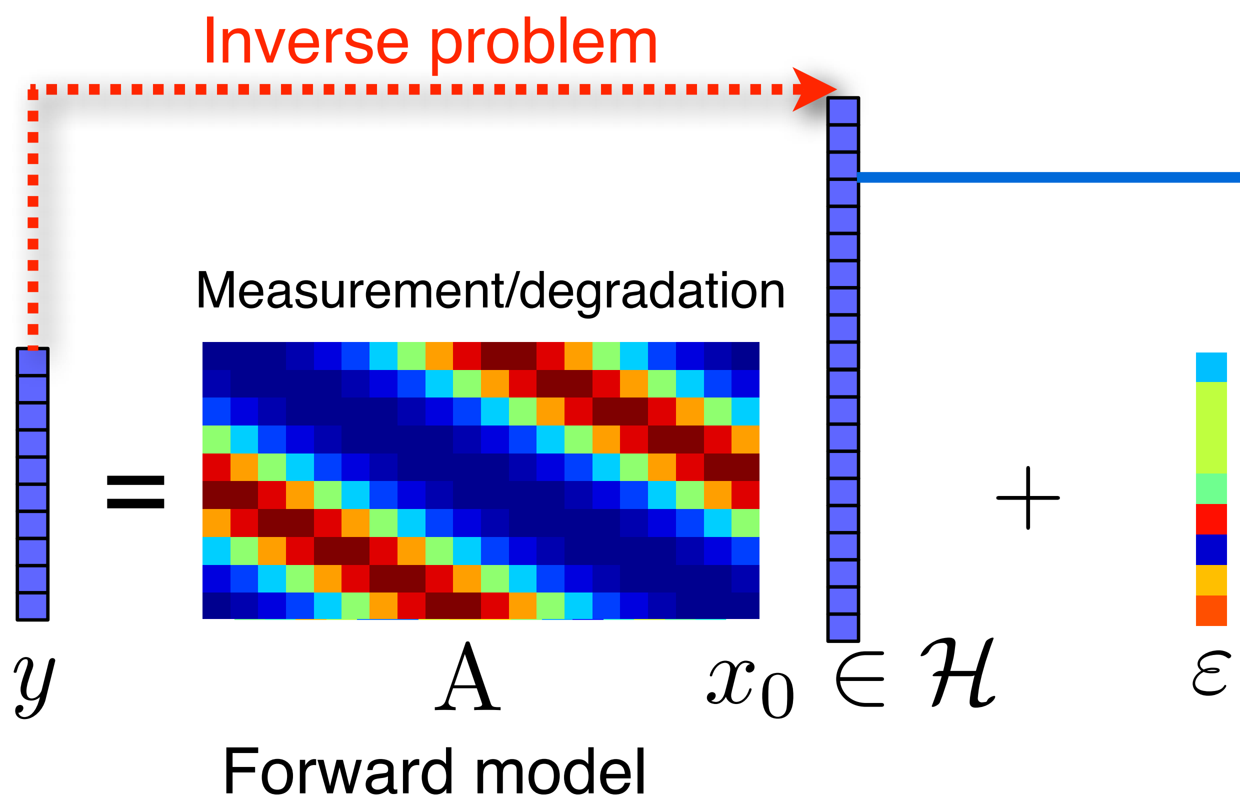
Outline

- Introduction.
- Non-smooth convex optimization.
  - Elements of convex analysis.
  - Elements of duality.
  - Optimality conditions.
- Proximal framework and operator splitting.
  - Proximal calculus.
  - Monotone operator splitting.
  - Sum of two functions.
  - Generalization to more than two functions.
- Take-away messages.

Today's lecture is about ...

- Non-smooth convex optimization.
- Convex analysis.
- Monotone operator splitting: divide and conquer.
- Fenchel-Rockafellar duality: think primal, act dual.
- Fast algorithms for e.g. data sciences.
- Connections with AJ lecture series.

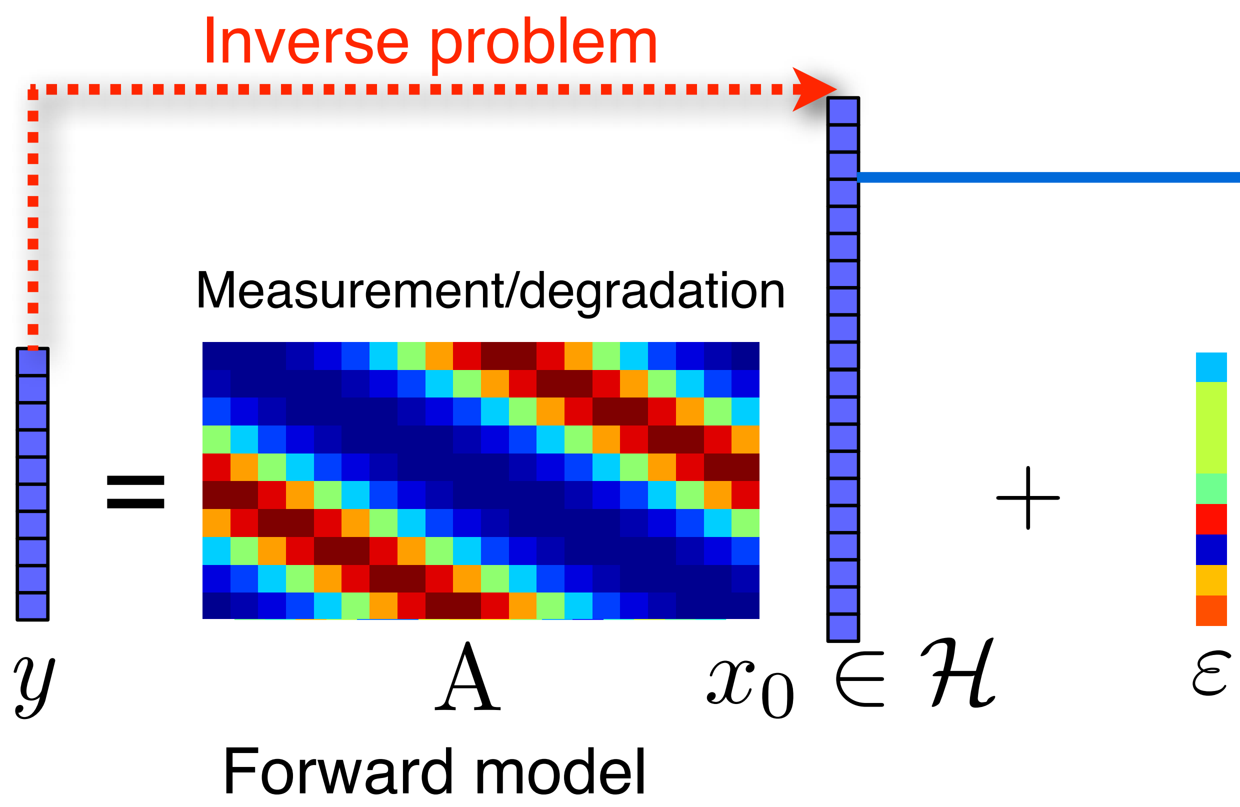
Regularized inverse problems



Prior knowledge
(regularization, constraints)

x_0 typically lives
in a low-dimensional manifold

Regularized inverse problems

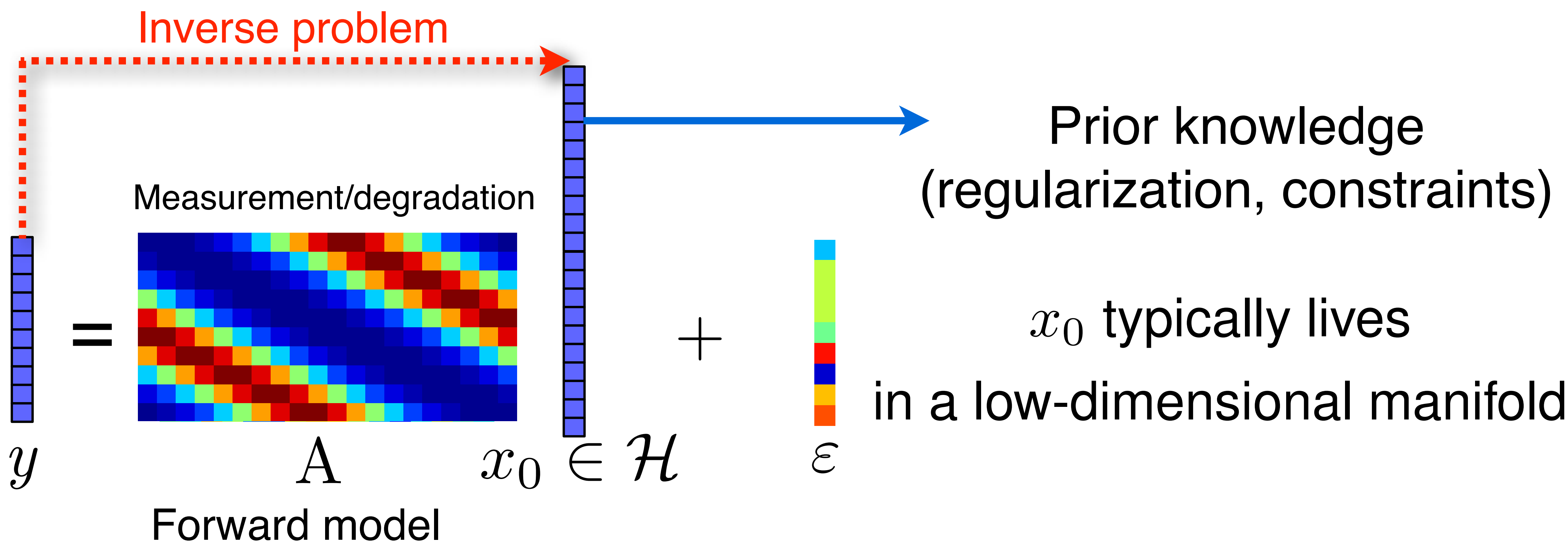


Prior knowledge
(regularization, constraints)

x_0 typically lives
in a low-dimensional manifold

- Many applications in data sciences: signal/image processing, machine learning, statistics, etc..

Regularized inverse problems



- Many applications in data sciences: signal/image processing, machine learning, statistics, etc..
- Solve an inverse problem through regularization :

$$\min_{x \in \mathcal{H}} \underbrace{F(x)}_{\text{Data fidelity}} + \underbrace{R(x)}_{\text{Regularization, constraints}}$$

- R promotes objects living in the same manifold as x_0.

Outline

- Introduction.
- Non-smooth convex optimization.
  - Elements of convex analysis.
  - Elements of duality.
  - Optimality conditions.
- Proximal framework and operator splitting.
  - Proximal calculus.
  - Monotone operator splitting.
  - Sum of two functions.
  - Generalization to more than two functions.
- Take-away messages.

Elements of convex analysis

Notations

- $\mathcal{H}$ is a finite-dimensional Hilbert space (typically the real vector space $\mathbb{R}^N$) endowed with the inner product $\langle ., . \rangle$ and associated norm $\|.\|$.
- I is the identity operator on $\mathcal{H}$.
- The operator spectral norm of $\mathrm{A} : \mathcal{H}_1 \to \mathcal{H}_2$ is denoted $|||\mathrm{A}||| = \sup_{x \in \mathcal{H}_1} \frac{\|\mathrm{A}x\|}{\|x\|}$.
- $\|.\|_p \, , p \geq 1$ is the ℓ_p-norm with the usual adaptation for the case $p = +\infty$.
- $\mathcal{B}_p^\rho$ is the (convex compact) ℓ_p-ball, $p \geq 1$, centered at its origin 0 and of radius $\rho > 0$. $x + \mathcal{B}_p^\rho$ is the same ball centered at x.

Sets

Definition (Convex set) *A closed set $\mathcal{C} \subseteq \mathcal{H}$ is said to be convex if :*

$$\forall x, y \in \mathcal{C}, \quad 0 \leq \rho \leq 1 \Rightarrow \rho x + (1-\rho) y \in \mathcal{C}.$$

Definition (Cone) *A cone $\mathcal{C}$ is a set such that the "open" half line $\{tx : t > 0\}$ is entirely contained in $\mathcal{C}$ whenever $x \in \mathcal{C}$. In the usual geometrical representation, a cone has an apex ; here at 0.*

Property (Convex cone) *A cone $\mathcal{C}$ is convex $\iff$ $\mathcal{C} + \mathcal{C} \subset \mathcal{C}$.*

Proposition (Convexity-preserving operations)

- *Convexity is stable under intersection : if $\mathcal{C}_i, i \in \mathcal{I}$ are convex $\Rightarrow \cap_{i \in \mathcal{I}} \mathcal{C}_i$ is convex.*
- *Convexity is stable under Cartesian product, and the converse is true : $\mathcal{C}_i, i \in \mathcal{I}$ are convex $\iff$ $\mathcal{C}_1 \times \cdots \times \mathcal{C}_{|\mathcal{I}|}$ is convex.*
- *Convexity is stable under affine mappings : the image of a convex set under an affine map A is also convex (e.g. reflection, Minkowski sum).*
- *If a set is convex, so are its interior and its closure.*

Sets

Definition (Affine hull) *An affine combination of* $x_1 \cdots x_n \in \mathcal{H}$ *is an element* $\sum_{i=1}^n a_i x_i$, $\sum_{i=1}^n a_i = 1$. *All such affine combinations form an affine manifold of* $\mathcal{H}$. *The affine hull of a nonempty set* $\mathcal{C} \subset \mathcal{H}$ *is the smallest affine manifolds containing* $\mathcal{C}$, *or equivalently,*

$$\operatorname{aff}(\mathcal{C}) = \left\{ x \in \mathcal{H} \;\middle|\; \forall i,\ y_i \in \mathcal{C},\ x = \sum_{i=1}^n a_i y_i \quad \textit{and} \quad \sum_{i=1}^n a_i = 1 \right\}.$$

- The interior of a convex set is empty unless it is full dimensional.
- Let $\mathcal{C}$ be a sheet of paper. Its interior is empty in the surrounding $\mathbb{R}^3$ space, ... but not in the space $\mathbb{R}^2$ of the table it is lying on.
- The concept of relative interior alleviates this ambiguity by defining the interior for a different topology : the one that equips its affine hull (which becomes a topological space in its own).
- In convex analysis and optimization, the topology of the whole space is of moderate interest, those relative to the affine hull are much richer.

Relative topology

Definition (Relative interior)

The relative interior $\mathrm{ri}(\mathcal{C})$ *of a convex set* $\mathcal{C} \subset \mathcal{H}$ *is the interior of* $\mathcal{C}$ *for the topology relative to its affine full, i.e.* $x \in \mathrm{ri}(\mathcal{C})$ *if and only if :*

$$x \in \mathrm{aff}(\mathcal{C}) \quad \textit{and} \quad \exists \rho > 0 \quad \text{s.t.} \quad (\mathrm{aff}(\mathcal{C})) \cap \mathcal{B}^{\rho}_{\mathcal{H}}(x) \subset \mathcal{C}\ .$$

$\mathcal{C}$	$\mathrm{aff}(\mathcal{C})$	$\dim(\mathcal{C})$	$\mathrm{ri}(\mathcal{C})$
$\{x\}$	$\{x\}$	0	$\{x\}$
$[x, x']$	affine line generated by x and x'	1	(x, x')
Simplex $\mathcal{S}_N$ in $\mathbb{R}^N$	affine manifold of equation $\sum_{i=1}^{N} x_i = 1$	$N-1$	$\{x \in \mathcal{S}_N : x[i] > 0\}$
$\mathcal{B}^{\rho}_{2} \subset \mathbb{R}^N$	$\mathbb{R}^N$	N	$\mathrm{int}(\mathcal{B}^{\rho}_{2})$

Proposition (Properties of the relative interior)

- $\mathrm{ri}(\mathcal{C}) \subset \mathcal{C}$*, is convex and* $\dim(\mathrm{ri}(\mathcal{C})) = \dim(\mathcal{C})$*.*
- *Let* $x \in \mathrm{cl}(\mathcal{C})$ *and* $x' \in \mathrm{ri}(\mathcal{C})$*, then* $(x, x'] \in \mathrm{ri}(\mathcal{C})$*.*
- *Consequently, the convex sets* $\mathcal{C}$*,* $\mathrm{ri}(\mathcal{C})$ *and* $\mathrm{cl}(\mathcal{C})$ *have the same affine hull, the same relative interior and the same closure.*
- *The relative topology fits well with convexity preserving operations. Let* $\mathcal{C}_i, i = 1, \cdots, n$ *be convex sets.*
  - *If* $\cap_i \mathrm{ri}(\mathcal{C}_i) \neq \emptyset \Rightarrow \cap_i \mathrm{ri}(\mathcal{C}_i) = \mathrm{ri}(\cap_i \mathcal{C}_i)$*.*
  - $\mathrm{ri}(\mathcal{C}_1) \times \cdots \times \mathrm{ri}(\mathcal{C}_n) = \mathrm{ri}(\mathcal{C}_1 \times \cdots \times \mathcal{C}_n)$*.*
  - *Let* A *be an affine map, then* $\mathrm{ri}(\mathrm{A}\mathcal{C}) = \mathrm{A}(\mathrm{ri}(\mathcal{C}))$*.*
- $0 \in \mathrm{ri}(\mathcal{C}_1 - \mathcal{C}_2) \iff \mathrm{ri}(\mathcal{C}_1) \cap \mathrm{ri}(\mathcal{C}_2) \neq \emptyset$*.*

Functions

Definition (Domain of a function) *The domain* $\mathrm{dom}(F)$ *of a function* $F : \mathcal{H} \to \mathbb{R}$ *is* $\mathrm{dom}(F) = \{x \in \mathcal{H} : F(x) < +\infty\}$.

Definition (Proper function) *A function is proper if* $\mathrm{dom}(F) \neq \emptyset$.

Definition (Epigraph, level set, sublevel sets) *The epigraph* $\mathrm{epi}(F)$ *of a function* $F : \mathcal{H} \to \mathbb{R}$ *is* $\mathrm{epi}(F) = \{(x, t) \in \mathcal{H} \times \mathbb{R} : F(x) \leq t\}$. *The level set of* F *at* t_0 *is* $\mathrm{lev}_{t_0}(F) = \{x \in \mathcal{H} : F(x) = t_0\}$. *The sublevel sets at* t_0 *is* $\cup_{t \leq t_0} \mathrm{lev}_t(F)$.

Definition (Coercivity) F *is (weakly- or 0-)coercive if* $\lim_{\|x\| \to \infty} F(x) = +\infty$.

Functions

Definition (Convex function) *A function $F : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is convex if*

$$\forall x, y \in \mathcal{H}, \quad 0 < \rho < 1, \quad F(\rho x + (1-\rho)y) \leq \rho F(x) + (1-\rho)F(y) \; .$$

It is strictly convex if the inequality is strict for $x \neq y$.

Definition (Lower semicontinuity) *We say that a real-valued function f is lower semi-continuous (lsc) if $\liminf_{x \to x_0} f(x) \geq f(x_0)$. It is lsc on $\mathcal{C} \subset \mathcal{H}$ if it is lsc at each of its points.*

Proposition *Let $F : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$. F is lsc $\iff$ its epigraph is closed $\iff$ its sublevel sets at t are closed for all $t \in \mathbb{R}$.*

Lower semi-continuity is weaker than continuity, and plays an important role for existence of solutions in minimization problems over a compact set (by closedness of its epigraph).

Notation $\Gamma_0(\mathcal{H})$ is the class of all proper lsc convex functions from $\mathcal{H}$ to $\mathbb{R} \cup \{+\infty\}$.

Properties of convex functions

Proposition (Properties of closed convex functions)

- *A function $F \in \Gamma_0(\mathcal{H})$ is (strictly) convex if and only if its epigraph is a (strictly) convex set.*
- *It is strongly convex with modulus c if $F - c/2 \|\cdot\|^2$ is convex.*
- *Any $F \in \Gamma_0(\mathcal{H})$ is minorized by some affine function : $F(y) \geq F(x) + \langle u, y - x \rangle, \forall x \in \mathrm{ri}(\mathrm{dom}(F)), \forall y \in \mathcal{H}$.*
- *Convexity and closedness of functions in $\Gamma_0(\mathcal{H})$ are preserved under :*
  - *positive combinations ;*
  - *pointwise supremum ;*
  - *(Legendre-Fenchel) conjugacy (see hereafter) ;*
  - *pre-composition by an affine mapping A such that $\mathrm{Im}(\mathrm{A}) \cap \mathrm{dom}(F) \neq \emptyset$;*
  - *post-composition $G \circ F$ with an increasing convex function $G \in \Gamma_0(\mathbb{R})$ if $\exists x \in \mathcal{H} \quad \text{s.t.} \quad F(x) \in \mathrm{dom}(G)$ and $G(+\infty) := +\infty$.*

Properties of convex functions

Theorem (Continuity properties) *Let F a convex function on $\mathbb{R}^N$.*

- *If $\mathcal{C}$ is a compact subset of $\mathrm{ri}(\mathrm{dom}(F))$, then F is continuous on the relative interior of its domain. It is moreover locally Lipschitz-continuous on this relative interior.*
- *If F is (uniformly) Lipschitz on a nonempty convex subset $\mathcal{C}$, it has a convex Lipschitz extension (with the same Lipschitz constant) on the whole space, that coincides with it on $\mathcal{C}$.*
- *Convex functions converging pointwise to some function F do converge uniformly on each compact subset of $\mathrm{ri}(\mathrm{dom}(F))$, and F is convex.*

Theorem (First-order properties) *Let F a convex function on $\mathbb{R}^N$.*

- *F is differentiable almost everywhere, i.e. the subset of $\mathrm{int}(\mathrm{dom}(F))$ where F is not differentiable is of zero Lebesgue measure.*
- *F differentiable on an open convex set $\mathcal{O} \iff F \in C^{1,1}(\mathcal{O})$.*

Theorem (Second-order properties [A.D. Alexandrov]) *Let F a convex function. For all $x \in \mathrm{int}(\mathrm{dom}(F))$ except on a set of zero Lebesgue measure, F is differentiable at x and there exists a symmetric positive semi-definite operator $\mathrm{D}^2F(x)$ such that for all $d \in \mathbb{R}^N$*

$$F(x+d) = F(x) + \langle \nabla F(x), d\rangle + \frac{1}{2}\left\langle \mathrm{D}^2F(x)d, d\right\rangle + o(\|d\|^2)\ .$$

Indicator and support functions

Definition (Indicator function) *Let $\mathcal{C}$ a nonempty subset of $\mathcal{H}$. The indicator function $\imath_{\mathcal{C}}$ of $\mathcal{C}$ is*

$$\imath_{\mathcal{C}}(x) = \begin{cases} 0, & \text{if } x \in \mathcal{C}\ , \\ +\infty, & \text{otherwise.} \end{cases}$$

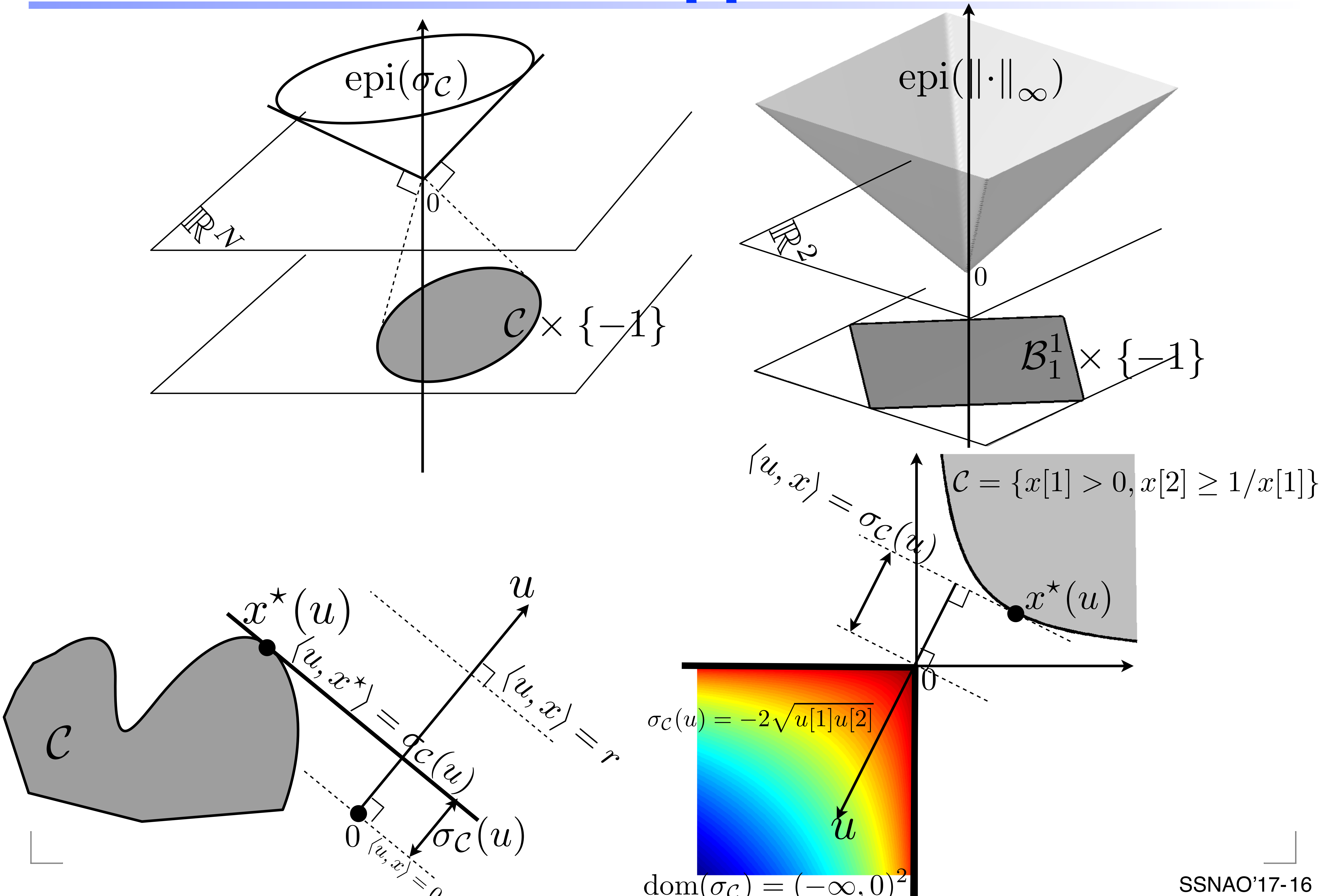
$\mathrm{dom}(\imath_{\mathcal{C}}) = \mathcal{C}$ *and* $\mathrm{epi}(\imath_{\mathcal{C}}) = \mathcal{C} \times \mathbb{R}^+$.

Definition (Support function) *Let $\mathcal{C}$ a nonempty subset of $\mathcal{H}$. Its support function is $\sigma_{\mathcal{C}}(u) = \sup\{\langle u, x\rangle : x \in \mathcal{C}\}, \forall u \in \mathcal{H}$; i.e. the supremum of the linear functions minorizing it.*

Proposition *$\sigma_{\mathcal{C}}$ is a closed convex function for any nonempty subset $\mathcal{C}$. It is sublinear ; i.e. positively homogeneous and subadditive, and is finite everywhere if $\mathcal{C}$ is bounded. Moreover, if $\mathcal{C}_1$ and $\mathcal{C}_2$ are nonempty closed convex sets, then $\mathcal{C}_1 \subset \mathcal{C}_2 \iff \sigma_{\mathcal{C}_1}(u) \leq \sigma_{\mathcal{C}_2}(u), \forall u \in \mathcal{H}$.*

Lemma *Any ℓ_p-norm is the support function of the unit ball $\mathcal{B}_q^1$ of the dual norm ℓ_q, where $1/p + 1/q = 1$.*

Indicator and support functions

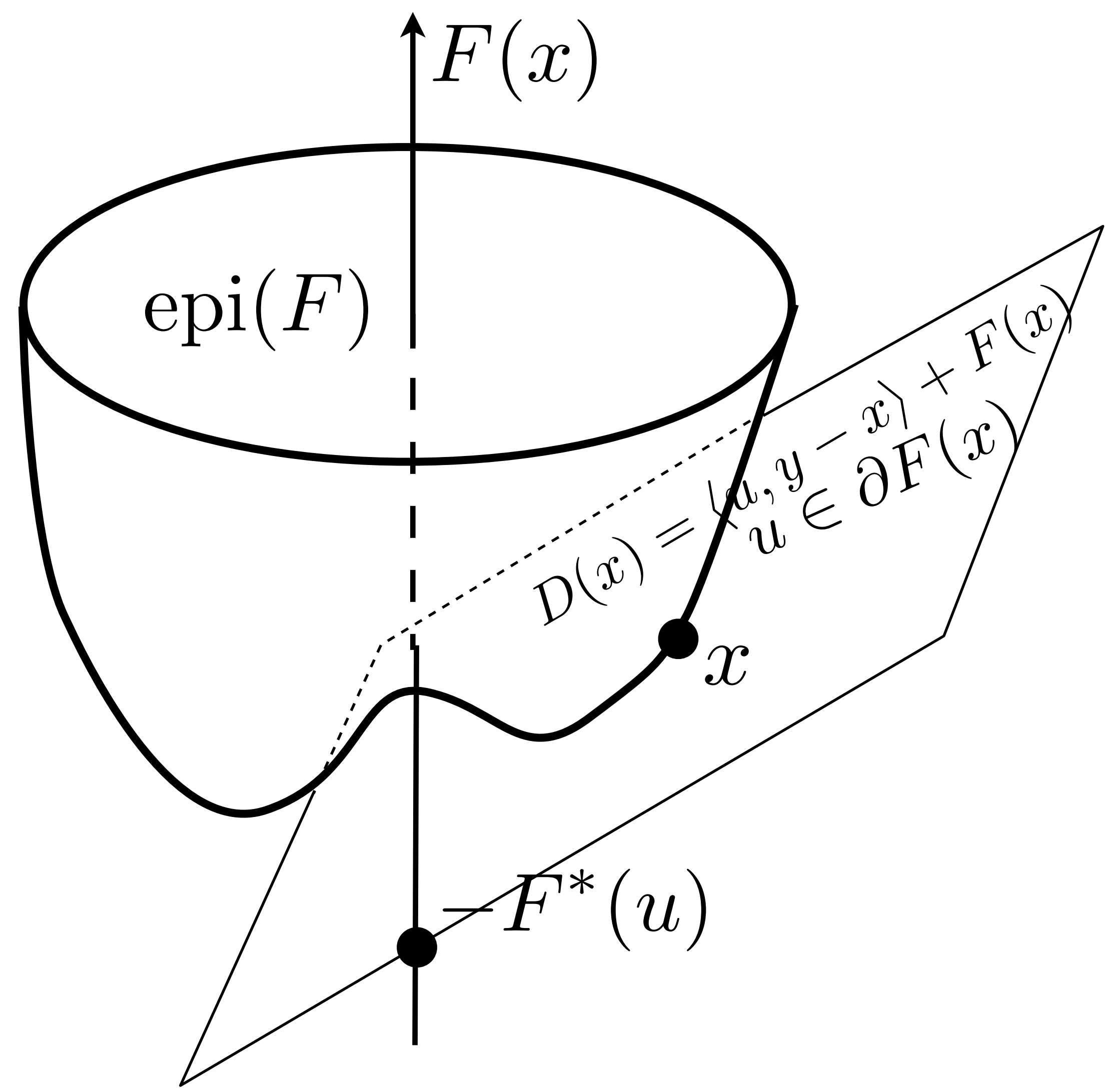


Conjugacy

Definition (Conjugate) *Let $F : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ having a minorizing affine function. The conjugate or Legendre-Fenchel transform of F is the function F^* defined by*

$$F^*(u) = \sup_{x \in \mathrm{dom}(F)} \langle u, x \rangle - F(x) \ .$$

We obviously observe that $F^*(u) + F(x) \geq \langle u, x \rangle$ for all $(x, u) \in \mathrm{dom}(F) \times \mathcal{H}$ (Fenchel inequality).



Conjugacy: properties

Theorem *F^* is a closed convex function. We also have $F \in \Gamma_0(\mathcal{H}) \iff$ the bi-conjugate $F^{**} = F$.*

Theorem (Calculus rules)

- $(F(x)+t)^*(u) = F^*(u) - t$.
- $(F(tx))^*(u) = tF^*(u/t)$, $t > 0$.
- $(F \circ \mathrm{A})^* = F^* \circ \left(\mathrm{A}^{-1}\right)^*$ *if A is a linear invertible operator.*
- $(F(x-x_0))^*(u) = F^*(u) + \langle u, x_0 \rangle$.
- $F_1 \leq F_2 \Rightarrow F_1^* \geq F_2^*$.
- *Separability :* $\left(\sum_{i=1}^n F_i(x_i)\right)^* = \sum_{i=1}^n F_i^*(u_i)$*, where* $(x_1, \cdots, x_n) \in \mathcal{H}_1 \times \cdots \times \mathcal{H}_n$.
- *Pre-composition with an affine operator : let $F \in \Gamma_0(\mathcal{H})$ and $\mathrm{A} := \mathrm{A}_0 \cdot + b$, an affine operator. Assume that $\mathrm{A}(\mathcal{H}) \cap \mathrm{ri}(\mathrm{dom}(F)) \neq \emptyset$. Then for avery $u \in \mathrm{dom}((F \circ \mathrm{A}_0)^*)$, the following minimization problem has a solution :*

$$(F \circ \mathrm{A})^*(u) = \inf_v \left\{F^*(v) - \langle v, b \rangle : \mathrm{A}_0^* v = u\right\} .$$

- *Conjugate of a sum : assume $F_1, F_2 \in \Gamma_0(\mathcal{H})$ and their relative interiors of their domains have a nonempty intersection. Then*

$$(F_1 + F_2)^* = F_1^* \overset{+}{\vee} F_2^* .$$

Conjugacy: differentiability

Theorem (First-order differentiability) *Let $F \in \Gamma_0(\mathcal{H})$ be strictly convex. Then $\mathrm{int}(\mathrm{dom}(F^*)) \neq \emptyset$ and F^* is continuously differentiable on $\mathrm{int}(\mathrm{dom}(F^*))$. Conversely, if $F \in \Gamma_0(\mathcal{H})$ is differentiable on $\mathrm{int}(\mathrm{dom}(F))$, then F^* is strictly convex on each convex subset $\mathcal{C} \subset \nabla F(\mathrm{int}(\mathrm{dom}(F)))$.*

Theorem (Second-order differentiability) *Assume that F is strongly convex on $\mathcal{H}$ with modulus c. Then F^* has full domain and a $1/c$-Lipschitz continuous gradient. Conversely, if $F \in \Gamma_0(\mathcal{H})$ has $1/c$-Lipschitz continuous gradient on $\mathcal{H}$, then F^* is strongly convex with modulus c on each convex subset of $\mathrm{dom}(\partial F^*)$.*

Conjugacy: examples

- The conjugate of the indicator function of a nonempty closed convex set is its support.
- Quadratic function : $F(x) := \frac{1}{2}\langle \mathrm{A}x, x\rangle + \langle b, x\rangle$, $\mathrm{A} \in \mathbb{R}^{n\times n} \succ 0$ and symmetric. $F^*(u) = \frac{1}{2}\left\langle u - b, \mathrm{A}^{-1}(u-b)\right\rangle$. If A is only semidefinite positive, we have $F^*(\mathrm{A}x + b) = \frac{1}{2}\langle x, \mathrm{A}x\rangle$.
- The conjugate of the directional derivative at x is the indicator of the subdifferential.
- Many other examples exploiting calculus rules in classical convex analysis monographs (see bibliography at the end).

Infimal convolution

Definition (Infimal convolution) *Let F_1 and F_2 two functions from $\mathcal{H}$ to $\mathbb{R} \cup \{+\infty\}$. Their infimal convolution is the function from $\mathcal{H}$ to $\mathbb{R} \cup \{\pm\infty\}$ defined by :*

$$(F_1 \overset{+}{\vee} F_2)(x) = \inf \{F_1(x_1) + F_2(x_2) : x_1 + x_2 = x\} = \inf_{y \in \mathcal{H}} F_1(y) + F_2(x - y) .$$

It is called exact at $x = \bar{x}_1 + \bar{x}_2$ if the infimum is attained at (non-necessarily unique) $(\bar{x}_1, \bar{x}_2)$.

Infimal convolution appears as a "convolution of infinite order" combined with exponentiation (in fact in a different algebra).

Proposition *Let F_1 and F_2 be convex functions.*

- *If F_1 and F_2 have a common affine minorant, then their inf-convolution is also convex.*
- *Inf-convolution of F_1 and F_2 is convex $\iff$ their strict epigraphs add up to the strict epigraph of their inf-convolution.*

Infimal convolution: properties

Theorem (Conjugate of an infimal convolution) *Let F_1 and F_2 be two proper functions (non-necessarily convex), such that the domain of their conjugates have a nonempty intersection, then*

$$(F_1 \overset{+}{\vee} F_2)^* = F_1^* + F_2^* \ .$$

In words, the Legendre-Fenchel conjugate acts as the Fourier transform in the $(\max, +)$ algebra.

Property

- *Domain :* $\mathrm{dom}(F_1 \overset{+}{\vee} F_2) = \mathrm{dom}(F_1) + \mathrm{dom}(F_2)$.
- *Inf-convolution is commutative, associative, its neutral element in $\Gamma_0(\mathcal{H})$ is $\imath_{\{0\}}$, and preserves the order.*

Example

- *Distance function : let $\mathcal{C}$ be a nonempty convex subset of $\mathcal{H}$, and $\|\cdot\|$ an arbitrary norm. Then the distance function to $\mathcal{C}$: $d_{\mathcal{C}} = \imath_{\mathcal{C}} \overset{+}{\vee} \|\cdot\|$.*
- *Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be two nonempty convex subsets, then $\imath_{\mathcal{C}_1} \overset{+}{\vee} \imath_{\mathcal{C}_2} = \imath_{\mathcal{C}_1 + \mathcal{C}_2}$.*
- *Moreau envelope : the function ${}^{\rho}F(x) = \inf_{z \in \mathcal{H}} \frac{1}{2\rho} \|x - z\|^2 + F(z) = F \overset{+}{\vee} \frac{1}{2\rho} \|\cdot\|^2$ for $0 < \rho < +\infty$ will be called the Moreau envelope of index ρ of F.*

Subdifferential

Definition (Directional derivative) *A function F admits a one-sided directional derivative at x in the direction d if*

$$F'(x,d) = \lim_{t \downarrow 0} \frac{F(x+td) - F(x)}{t} = \inf_{t>0} \frac{F(x+td) - F(x)}{t}$$

exists with values in $[-\infty, \infty]$. It is two-sided if and only if $F'(x,-d)$ exists and $F'(x,-d) = -F'(x,d)$.

Definition (Subdifferential I) *The subdifferential of a function $F \in \Gamma_0(\mathcal{H})$ at $x \in \mathcal{H}$ is the set-valued map $\partial F : \mathcal{H} \to 2^{\mathcal{H}}$*

$$\partial F(x) = \{u \in \mathcal{H} : \forall z \in \mathcal{H}, F(z) \geq F(x) + \langle u, z - x \rangle\} \ ,$$

i.e., the set of slopes of affine functions minorizing F at x. An element u of $\partial F(x)$ is called a subgradient. The subdifferential of the indicator function of a closed convex set $\mathcal{C}$ is the normal cone of $\mathcal{C}$ at x :

$$\mathcal{N}_{\mathcal{C}}(x) = \{u \in \mathcal{H} : \langle u, x - z \rangle \geq 0, \forall z \in \mathcal{C}\} \ .$$

Definition (Subdifferential II) *The subdifferential of $f \in \Gamma_0(\mathcal{H})$ at $x \in \mathcal{H}$ if the nonempty compact convex set whose support function is the directional derivative $F'(x,d)$.*

$$\partial F(x) = \{d \in \mathcal{H} : F'(x,d) \geq \langle u, d \rangle\} \ .$$

Subdifferential properties

Theorem (Properties of the subdifferential) *Let F be a convex function.*

- *For fixed x, $F'(x,d)$ is finite sublinear (hence convex in d).*
- *Monotonicity :*
  - *A function F is convex on a convex set $\mathcal{C}$ $\iff$ $\langle u_1 - u_2, x_1 - x_2 \rangle \geq 0, \ \forall u_i \in \partial F(x_i), x_i \in \mathcal{C}, \ i = 1,2$ (i.e. ∂F is monotone).*
  - *F is strictly convex on a convex set $\mathcal{C}$ $\iff$ the subdifferential inequality becomes strict for $x_1 \neq x_2 \in \mathcal{C}$ $\iff$ $\langle u_1 - u_2, x_1 - x_2 \rangle > 0$ (i.e. ∂F is strictly monotone).*
  - *F is strongly convex with modulus $c > 0$ $\iff$ $\langle u_1 - u_2, x_1 - x_2 \rangle \geq c \left\| x_1 - x_2 \right\|^2$ $\iff$ $F(x_2) \geq F(x_1) + \langle u, x_2 - x_1 \rangle + \frac{c}{2} \left\| x_2 - x_1 \right\|^2, \forall x_2 \in \mathcal{H}$ (i.e. ∂F is strongly monotone).*
- *Continuity :*
  - *$\partial F(x) = \{\nabla F(x)\}$ almost everywhere, except on a set of (Lebesgue) measure zero (kinks).*
  - *If F is (Gâteaux) differentiable at x, its only subgradient at x is its gradient $\nabla F(x)$. Conversely, if $\partial F(x) = \{u\}$, then F is (Fréchet) differentiable at x, with $\nabla F(x) = u$.*
- *The subdifferential can be defined in terms of F and its conjugate F^*,*

$$u \in \partial F(x) \iff F(x) + F^*(u) = \langle x, u \rangle \iff x \in \partial F^*(u) \ .$$

Subdifferential calculus

Theorem (Calculus rules with subdifferentials) *Let all considered functions be proper convex.*

- *Positive linear combinations : if* $\bigcap_i \mathrm{ri}(\mathrm{dom}(f_i)) \neq \emptyset$*, then* $\partial(\sum_{i=1}^n \rho_i F_i)(x) = \sum_{i=1}^n \rho_i \partial F_i(x), \rho_i \geq 0, i = 1, \cdots, n$.
- *Pre-composition with an affine mapping : let* A *be an affine mapping :* $\mathrm{A} := \mathrm{A}_0 \cdot + b$, A_0 *is linear, such that* $\mathrm{Im}(\mathrm{A}) \cap \mathrm{ri}(\mathrm{dom}(\mathrm{A}_0)) \neq \emptyset$*. Then* $\partial(F \circ \mathrm{A})(x) = \mathrm{A}_0^* \partial F(\mathrm{A}x)$.
- *Pointwise supremum :* $F(x) := \sup_{i \in \mathcal{I}} F_i(x)$*, where* $\mathcal{I}$ *is compact. Let* $\mathcal{I}(x) = \{i : F(x) = F_i(x)\}$*. Then,*

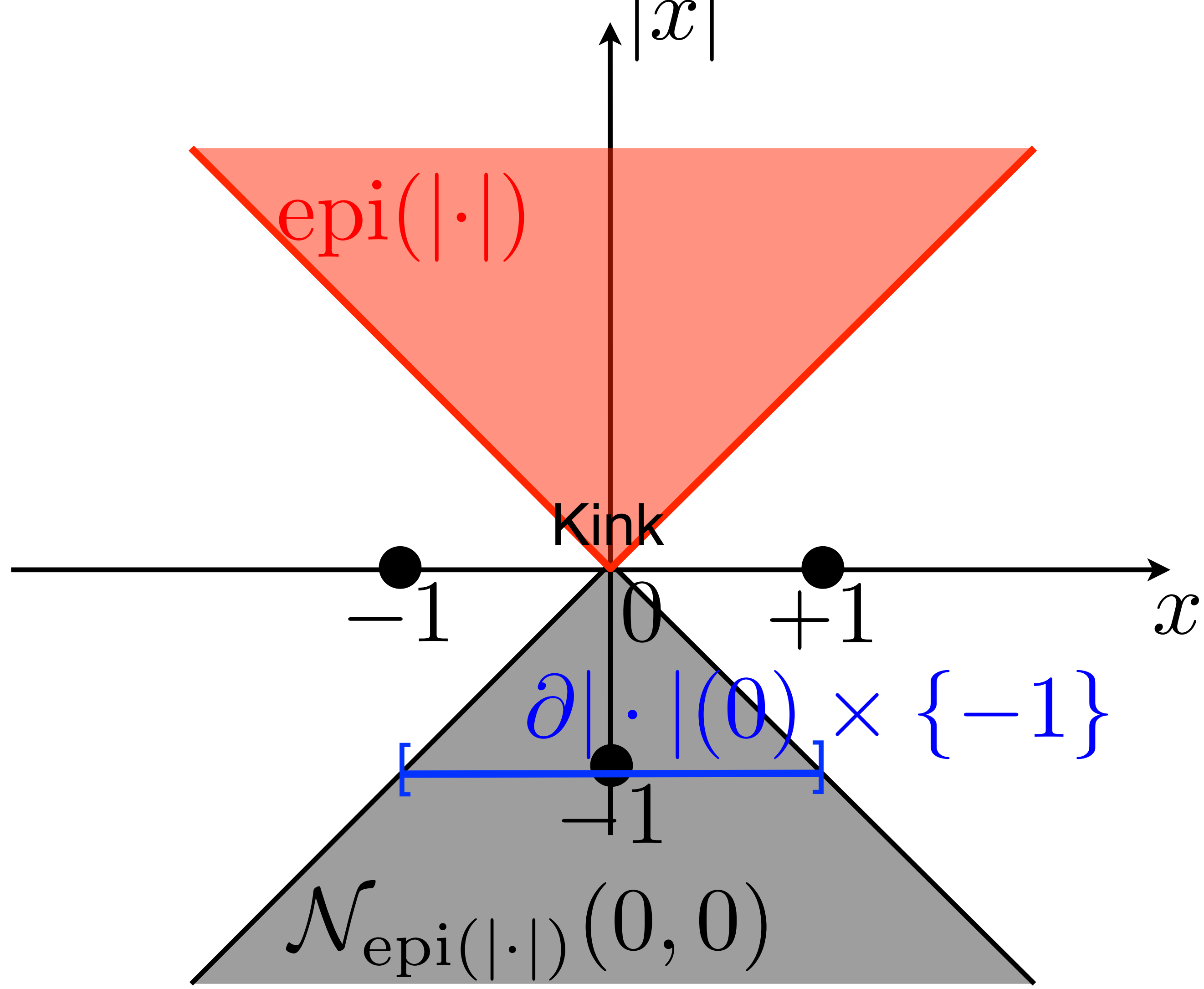
$$
\begin{aligned}
\partial F(x) &= \left\{ \sum_{i \in \mathcal{I}(x)} \rho_i \partial F_i(x), \rho_i \geq 0 \text{ for all } i \in \mathcal{I}(x), \sum_{i \in \mathcal{I}(x)} \rho_i = 1 \right\} \\
&= \mathrm{convhull} \left\{ \cup_{i \in \mathcal{I}(x)} \partial F_i(x) \right\} .
\end{aligned}
$$

Subdifferential: geometric interpretation

Theorem (Subdifferential III) *Let $F \in \Gamma_0(\mathcal{H})$. A point u is a subgradient of F at x if and only if $(u, -1)$ is normal to $\mathrm{epi}(F)$ at $(x, F(x))$; i.e.*

$$\mathcal{N}_{\mathrm{epi}(F)}(x, F(x)) = \lambda(\partial F(x) \times \{-1\}), \lambda \geq 0 \ .$$

In other words, the intersection of the normal cone of $\mathrm{epi}(F)$ and $\mathcal{H}$ at level -1 is just the subdifferential $\partial F(x)$ shifted vertically in $\mathcal{H} \times \mathbb{R}$ by -1.



Outline

- Introduction.
- Non-smooth convex optimization.
  - Elements of convex analysis.
  - Elements of duality.
  - Optimality conditions.
- Proximal framework and operator splitting.
  - Proximal calculus.
  - Monotone operator splitting.
  - Sum of two functions.
  - Generalization to more than two functions.
- Take-away messages.

Fenchel-Rockafellar duality

The duality formula to be stated shortly plays an important role in dualizing optimization problems (e.g. proximity operator calculus, ADMM for the augmented-Lagrangian method, and many, many other situations).

Theorem (Fenchel-Rockafellar duality) *Let* $F \in \Gamma_0(\mathcal{H})$ *and* $G \in \Gamma_0(\mathcal{K})$*, and* $\mathrm{A} := \mathrm{A}_0 \cdot -b : \mathcal{H} \to \mathcal{K}$ *be a bounded affine operator, and* $\mathcal{H}$ *and* $\mathcal{K}$ *are finite-dimensional real Hilbert space (as we supposed from the beginning). Suppose that* $0 \in \mathrm{ri}(\mathrm{dom}(G)) - \mathrm{A}\left(\mathrm{ri}(\mathrm{dom}(F))\right)$*. Then*

$$\inf_{x \in \mathcal{H}} F(x) + G \circ \mathrm{A}(x) = -\min_{u \in \mathcal{K}} F^*(-\mathrm{A}_0^* u) + G^*(u) + \langle u, b \rangle \ ,$$

with the relashionships between $x^\star$ *and* $u^\star$*, respectively the solutions of the primal and dual problems*

$$\begin{aligned} F(x^\star) + F^*(-\mathrm{A}_0^* u^\star) &= \langle -\mathrm{A}_0^* u^\star, x^\star \rangle \,, \\ G(\mathrm{A}x^\star) + G^*(u^\star) &= \langle u^\star, \mathrm{A}x^\star \rangle \,, \end{aligned}$$

or equivalently $(x^\star, u^\star)$ *are the so-called Kuhn-Tucker pairs :*

$$\begin{aligned} x^\star \in \partial F^*(-\mathrm{A}_0^* u^\star) \quad &\text{and} \quad u^\star \in \partial G(\mathrm{A}x^\star) \ , \\ -\mathrm{A}_0^* u^\star \in \partial F(x^\star) \quad &\text{and} \quad \mathrm{A}x^\star \in \partial G^*(u^\star) \ . \end{aligned}$$

From Fenchel-Rockafellar to Lagrange

$$(\mathsf{P}) : \inf_{x \in \mathcal{H}} F(x) + G \circ \mathrm{A}(x) \ ,$$

is equivalent to

$$\inf_{(x,z) \in \mathcal{H} \times \mathcal{K}} F(x) + G(z) \quad \text{s.t.} \quad z = \mathrm{A}x \ .$$

This is a minimization problem in $\mathcal{H} \times \mathcal{K}$ with equality constraint-values in $\mathcal{K}$, which lends itself to Lagrange duality : form the Lagrangian $L(x, z, u)$ with the dual variable u in $\mathcal{K}$:

$$L(x, z, u) = F(x) + G(z) + \langle u, \mathrm{A}x - z \rangle \ .$$

Th associated closed convex dual function is :

$$H(u) = \inf_{x,z} L(x, z, u) = -\sup_{x,z} \langle u, b \rangle + (\langle u, -\mathrm{A}_0 x \rangle - F(x)) + (\langle u, z \rangle - G(z)) \ .$$

By conjugacy calculus rules we obtain,

$$-H(u) = F^*(-\mathrm{A}_0^* u) + G^*(u) + \langle u, b \rangle \ .$$

The (Lagrange) dual problem is then :

$$(\mathsf{Q}) : \max_{u \in \mathcal{K}} H(u) = -\min_{u \in \mathcal{K}} F^*(-\mathrm{A}_0 u) + G^*(u) + \langle u, b \rangle \ .$$

Fenchel-Rockafellar duality: example

Proposition *Let* $\mathrm{A} : \mathbb{R}^n \to \mathbb{R}^m$ *be a linear operator with a nonempty range. Then the following primal and dual problems are equivalent :*

$$
\begin{aligned}
(\mathsf{P}) : \quad & \inf_{x\in\mathbb{R}^n} \quad \frac{1}{2}\left\|y - \mathrm{A}x\right\|_2^2 + \lambda \left\|x\right\|_1 \\
(\mathsf{Q}) : \quad & \min_{u\in\mathbb{R}^m} \quad \left\|y - u\right\|_2 \quad \text{s.t.} \quad \left\|\mathrm{A}^{\mathrm{T}}u\right\|_\infty \leq \lambda \, .
\end{aligned}
$$

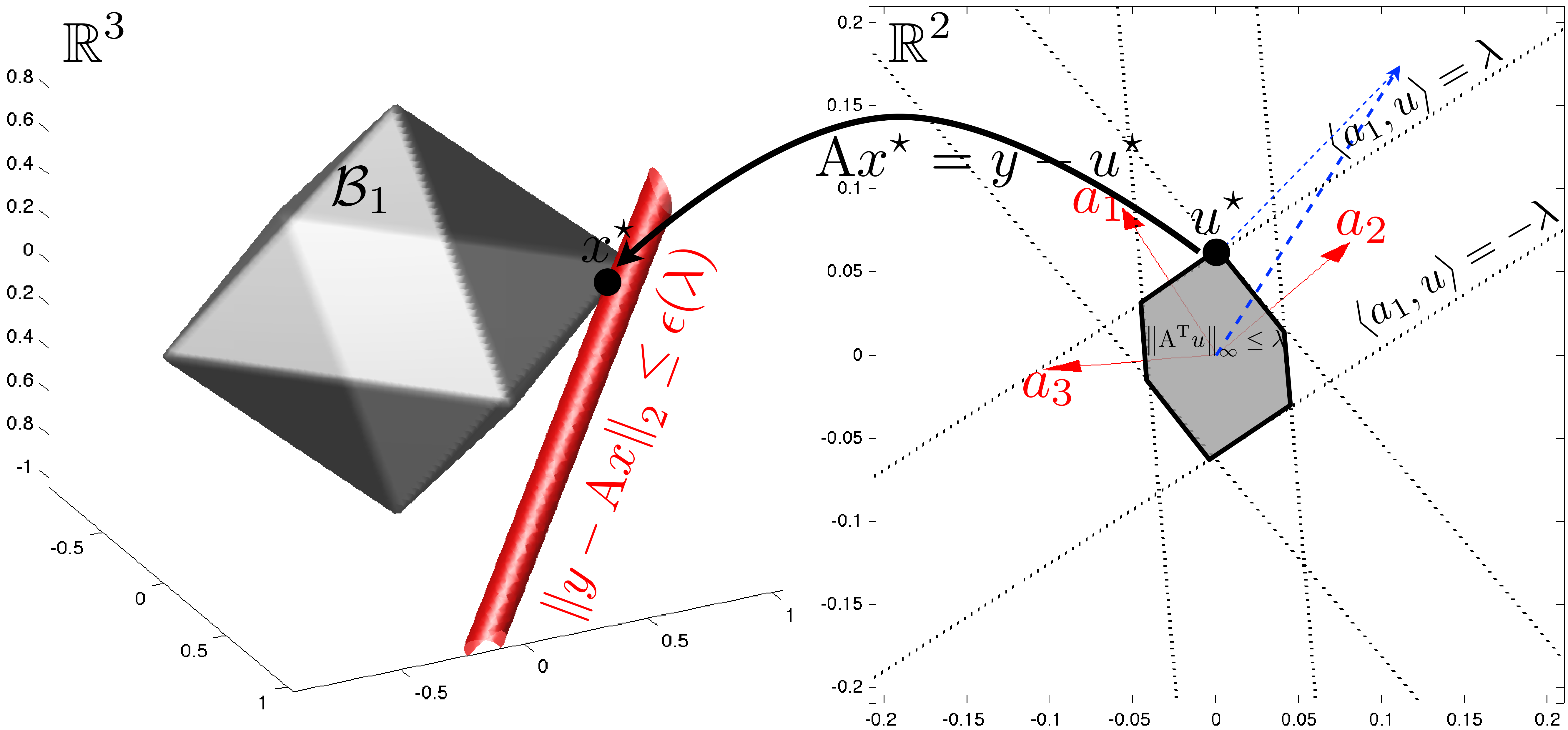
The primal solution to (P) *is related to the dual one (i.e. that of* (Q)*) as* $\mathrm{A}x^\star = y - u^\star$*.*

Proof: Use Fenchel-Rockafellar duality lemma, conjugacy calculus rules (quadratic function, norm, translation, scaling), and continuity properties of the conjugate. ■

Fenchel-Rockafellar duality: example

$$(\mathsf{P}): \quad \inf_{x\in\mathbb{R}^n} \quad \frac{1}{2}\|y - \mathrm{A}x\|_2^2 + \lambda\|x\|_1$$

$$(\mathsf{Q}): \quad \min_{u\in\mathbb{R}^m} \quad \|y-u\|_2 \quad \text{s.t.} \quad \left\|\mathrm{A}^{\mathrm{T}}u\right\|_\infty \le \lambda\,.$$



Outline

- Introduction.
- Non-smooth convex optimization.
  - Elements of convex analysis.
  - Elements of duality.
  - Optimality conditions.
- Proximal framework and operator splitting.
  - Proximal calculus.
  - Monotone operator splitting.
  - Sum of two functions.
  - Generalization to more than two functions.
- Take-away messages.

Optimality conditions

$$\text{(P)} \min_{x \in \mathcal{H}} F(x), \quad F \in \Gamma_0(\mathcal{H}).$$

Theorem (Minimality conditions) *Assume that the set of minimizers is nonempty, e.g. by coercivity. The following statements are equivalent :*

(i) $x^\star$ *is a global minimizer of* $F \in \Gamma_0(\mathcal{H})$ *over* $\mathcal{H}$ *;*

(ii) $0 \in \partial F(x^\star)$ *;*

(iii) $F'(x^\star, d) \geq 0$ *for all* d*.*

(iv) $x^\star$ *is a solution to the fixed point equation* $x = (\mathrm{I} + \mu \partial F)^{-1}(x)$*.*

- The fixed point equation in (iv) underlies the proximal iteration (or algorithm). Why ? Keep listening.
- $(\mathrm{I} + \mu \partial F)^{-1}$ is the resolvent associated to the subdifferential, see shortly.
- The above statements can be generalized to minimizers relative to a closed convex set (in terms of the normal and tangent cones), i.e. convex programming with nonsmooth objectives. But this path will deliberately not be pursued here because constraints are implicit in F.

Outline

- Introduction.
- Non-smooth convex optimization.
  - Elements of convex analysis.
  - Elements of duality.
  - Optimality conditions.
  - Sub-gradient descent.
- Proximal framework and operator splitting.
  - Proximal calculus.
  - Sum of two functions.
  - Generalization to more than two functions.
- Take-away messages.

Subgradient descent: the gist

$$\text{(P)} \min_{x \in \mathcal{H}} F(x), \quad F \in \Gamma_0(\mathcal{H}).$$

- Follow the footprints of (possibly projected) gradient descent for smooth optimization.
- Replace the gradient by a subgradient $u_k \in \partial F(x_k)$.
- However, serious difficulties : no line search is possible based on decreasing F, simply because
  - u_k may not be a descent direction (e.g. think of the ℓ_1-norm). Thus oscillations in the objective (non-monotonic behaviour) ;
  - u_k is so weak that the resulting sequence would not minimize F.

Subgradient descent scheme

Initialization : Choose a sequence of step sizes $(\mu_k)_{k\in\mathbb{N}}$, $\mu_k > 0$. Choose an initial $x_0 \in \mathrm{dom}(F)$ and obtain $u_0 \in \partial F(x_0)$. $k = 0$

Main iteration : Construct a sequence of iterates $(x_k)_{k\in\mathbb{N}}$ as follows :

repeat

$$x_{k+1} = x_k - \mu_k \frac{u_k}{\max(\|u_k\|, 1)} .$$

Get $u_{k+1} \in \partial F(x_{k+1})$;

$k \leftarrow k + 1$.

until *convergence*;

- How to choose μ_k ?

Subgradient descent: Convergence

Theorem (Global convergence of Subgradient Descent) *Let $F \in \Gamma_0(\mathcal{H})$ and apply the subgradient descent algorithm with a sequence of step sizes satisfying :*

$$\lim_{k\to\infty} \mu_k = 0 \quad \textit{and} \quad \sum_{k\in\mathbb{N}} \mu_k = +\infty \; .$$

Then $F(x_k) \to \inf_x F(x)$ and $x_k \to x^\star \in M^\star$, $x^\star$ not necessarily unique.

- Typical choices : $\mu_k = \frac{1}{(k+1)^p}, p \in (0,1]$ or $\mu_k = \frac{1}{(k+1)\log(k+1)}$.
- Not easy to choose in practice and some sequences lead to a very slow convergence.
- Some elaborated choices are possible in the literature, but extra information such as knowledge about the solution set is needed.
- The choice is even more complicated by floating-point computations : it is hard to satisfy simultaneously the two step size requirements accurately.
- the stopping rule is not convenient : u_k has no reason to tend to 0. Stopping rule when μ_k becomes very small (compared to the scale of the problem).

Subgradient descent: Convergence

Theorem (Complexity result) *Let F be nonsmooth convex function. Then, no iterative scheme to minimize F relying only on its first-order properties (i.e. F and ∂F) can achieve a better rate than $O(1/\sqrt{k})$ on the objective.*

- In other words, we need $O(1/\epsilon^2)$ to reach an ϵ-accurate solution on the objective.
- Other methods to circumvent these difficulties. Many of them exploit the structure of F to get more powerful provably convergent algorithms. This is what we are about to do.

Outline

- Introduction.
- Non-smooth convex optimization.
  - Elements of convex analysis.
  - Elements of duality.
  - Optimality conditions.
- Proximal framework and operator splitting.
  - Proximal calculus.
  - Monotone operator splitting.
  - Sum of two functions.
  - Generalization to more than two functions.
- Take-away messages.

Proximity operator

The notion of a proximity operator was introduced as a generalization in [J.-J. Moreau 1962] of convex projection operator.

Definition (Proximity operator) *Let $F \in \Gamma_0(\mathcal{H})$. Then, for every $x \in \mathcal{H}$, the function $z \mapsto \frac{1}{2}\|x - z\|^2 + F(z)$ achieves its infimum at a unique point denoted by $\operatorname{prox}_F x$. The uniquely-valued operator $\operatorname{prox}_F : \mathcal{H} \to \mathcal{H}$ thus defined is the proximity operator of F. It will be convenient to introduce the reflection operator $\operatorname{rprox}_F = 2\operatorname{prox}_F - \mathrm{I}$.*

Proximity operator: properties

Theorem (Some properties of the proximity operator) *Let $F \in \Gamma_0(\mathcal{H})$.*

- *Let $\forall x, z \in \mathcal{H}$, then*

$$p = \operatorname{prox}_F x \iff x - p \in \partial F(p) \; .$$

  Or equivalently, $\operatorname{prox}_F = (\mathrm{I} + \partial F)^{-1}$, prox_F is the resolvent of the subdifferential of F, a maximal monotone operator from $\mathcal{H} \to 2^{\mathcal{H}}$.

- *Continuity : the proximity operator is firmly nonexpansive. Hence its is nonexpansive and so is its reflection operator, and therefore they are both continuous on $\mathcal{H}$ into itself.*

Moreau envelope

Definition (Moreau envelope) *The function* ${}^{\rho}F(x) = \inf_{z\in\mathcal{H}} \frac{1}{2\rho}\|x-z\|^2 + F(z)$ *for* $0 < \rho < +\infty$ *is the Moreau envelope of index* ρ *of* F*.* ${}^{\rho}F$ *is also the infimal convolution of* F *with* $\frac{1}{2\rho}\|\cdot\|^2$*.*

Moreau envelope: properties

Lemma *Let $F \in \Gamma_0(\mathcal{H})$. Then its Moreau envelope ${}^\rho F$ is convex and Fréchet-differentiable with $1/\rho$-Lipschitz gradient*

$$\nabla {}^\rho F = (\mathrm{I} - \mathrm{prox}_{\rho F})/\rho.$$

Furthermore, its proximity operator is the convex combination

$$\mathrm{prox}_{{}^\rho F}(x) = \frac{\rho}{1+\rho} x + \frac{1}{1+\rho} \mathrm{prox}_{(1+\rho)F}(x) \ .$$

Because of the $C^{1,1}$-smoothness of ${}^\rho F$, the Moreau envelope is also known as the Moreau-Yosida regularization of F.

Lemma (Moreau identity) *Let $F \in \Gamma_0(\mathcal{H})$, then for any $x \in \mathcal{H}$*

$$\mathrm{prox}_{\rho F^*}(x) + \rho \, \mathrm{prox}_{F/\rho}(x/\rho) = x, \forall \, 0 < \rho < +\infty \ .$$

Corollary *Let $F \in \Gamma_0(\mathcal{H})$, then for any $x \in \mathcal{H}$*

$$\mathrm{prox}_{F^*} = \mathrm{I} - \mathrm{prox}_F \iff \mathrm{prox}_{F^*}(x) \in \partial F(\mathrm{prox}_F(x)) \ .$$

A detailed example

$$F(x) = |x|$$

A detailed example

$$F(x) = |x|$$

A detailed example

Outline

- Introduction.
- Non-smooth convex optimization.
  - Elements of convex analysis.
  - Elements of duality.
  - Optimality conditions.
- Proximal framework and operator splitting.
  - Proximal calculus.
  - Monotone operator splitting.
  - Sum of two functions.
  - Generalization to more than two functions.
- Take-away messages.

Proximal calculus

Proposition (Simple calculus rules) *Let $F \in \Gamma_0(\mathcal{H})$ and $x \in \mathcal{H}$.*

1. *Quadratic perturbation : let $G = F + \zeta \|.\|^2/2 + \langle ., u\rangle + \beta$, with $u \in \mathcal{H}$, $\zeta \in [0, +\infty)$ and $\beta \in \mathbb{R}$. then $\operatorname{prox}_G x = \operatorname{prox}_{F/(\zeta+1)}((x-u)/(\zeta+1))$.*
2. *Translation : let $G = F(. - z)$, with $z \in \mathcal{H}$. Then $\operatorname{prox}_G x = z + \operatorname{prox}_F(x - z)$.*
3. *Scaling : let $G = f(./\zeta)$, with $\zeta \in \mathbb{R} \setminus \{0\}$. Then $\operatorname{prox}_G x = \zeta \operatorname{prox}_{F/\zeta^2}(x/\zeta)$.*
4. *Reflexion : let $G : x \mapsto F(-x)$. Then $\operatorname{prox}_G x = -\operatorname{prox}_F(-x)$.*
5. *Separability : let $(F_i)_{1\leq i\leq n}$ a family of functions each in $\Gamma_0(\mathcal{H}_i)$, and $F : \mathcal{H} = \mathcal{H}_1 \times \cdots \times \mathcal{H}_n \to \mathbb{R}$ such that $F(\alpha) = \sum_{i=1}^n F_i(\alpha_i)$, $\alpha_i \in \mathcal{H}_i$. Then F is in $\Gamma_0(\mathcal{H})$ and $\operatorname{prox}_F = \{\operatorname{prox}_{F_i}\}_{1\leq i\leq n}$.*

Many others are available or can be calculated.

Proximity operator of $F \circ \mathrm{A}$

Lemma *Let $F \in \Gamma_0(\mathcal{K})$ and $\mathrm{A} = \mathrm{A}_0 \cdot -y$, where $\mathrm{A}_0 : \mathcal{H} \to \mathcal{K}$ is a bounded linear operator, and $\mathcal{H}$ and $\mathcal{K}$ are finite-dimensional.*

(i) If A_0 is a tight frame with constant c. Then

$$\mathrm{prox}_{F \circ \mathrm{A}}(x) = x + c^{-1}\mathrm{A}_0^* \left(\mathrm{prox}_{cF} - \mathrm{I}\right)\left(\mathrm{A}_0 x - y\right) .$$

(ii) If A_0 is a general frame with bounds c_1 and c_2. Let $\mu_k \in (0, 2/c_2)$. Define

$$\begin{aligned} u_{k+1} =& \mu_k \left(\mathrm{I} - \mathrm{prox}_{\mu_k^{-1} F}\right) \circ \left(\mu_k^{-1} u_k + \mathrm{A}\left(p_k\right)\right) \ , \\ p_{k+1} =& x - \mathrm{A}_0^* u_{k+1} \ . \end{aligned}$$

Then $p_k \to \mathrm{prox}_{F \circ \mathrm{A}}$ linearly.

(iii) If $c_1 = 0$ and $F \circ \mathrm{A} \in \Gamma_0(\mathcal{H})$ (typically if A is such that $\mathrm{ri}(\mathrm{dom}(F) \cap \mathrm{Im}(\mathrm{A})) \neq \emptyset$). Apply the above iteration with $\mu_k \in (0, 2/c_2)$. Then $p_k \to \mathrm{prox}_{F \circ \mathrm{A}}$ at the rate $O(1/k)$.

Proximity operator of $F \circ \mathrm{A}$

Lemma *Let $F \in \Gamma_0(\mathcal{K})$ and $\mathrm{A} = \mathrm{A}_0 \cdot -y$, where $\mathrm{A}_0 : \mathcal{H} \to \mathcal{K}$ is a bounded linear operator, and $\mathcal{H}$ and $\mathcal{K}$ are finite-dimensional.*

(i) If A_0 is a tight frame with constant c. Then

$$\operatorname{prox}_{F\circ\mathrm{A}}(x) = x + c^{-1}\mathrm{A}_0^* \left(\operatorname{prox}_{cF} - \mathrm{I}\right)\left(\mathrm{A}_0 x - y\right) .$$

(ii) If A_0 is a general frame with bounds c_1 and c_2. Let $\mu_k \in (0, 2/c_2)$. Define

$$\begin{aligned} u_{k+1} =& \mu_k \left(\mathrm{I} - \operatorname{prox}_{\mu_k^{-1}F}\right) \circ \left(\mu_k^{-1} u_k + \mathrm{A}\left(p_k\right)\right) \ , \\ p_{k+1} =& x - \mathrm{A}_0^* u_{k+1} \ . \end{aligned}$$

Then $p_k \to \operatorname{prox}_{F\circ\mathrm{A}}$ linearly.

(iii) If $c_1 = 0$ and $F\circ\mathrm{A} \in \Gamma_0(\mathcal{H})$ (typically if A is such that $\operatorname{ri}(\operatorname{dom}(F)\cap\operatorname{Im}(\mathrm{A})) \neq \emptyset$). Apply the above iteration with $\mu_k \in (0, 2/c_2)$. Then $p_k \to \operatorname{prox}_{F\circ\mathrm{A}}$ at the rate $O(1/k)$.

- Multi-step (e.g. inertial, see in the sequel) algorithms can be used as well (linear or $O(1/k^2)$ rate).
- Robustness to errors.

Proximity operator of $F_1 + F_2 \circ \mathrm{A}$

Lemma *Let $F_1 \in \Gamma_0(\mathcal{H})$ and $F_2 \in \Gamma_0(\mathcal{K})$, and $\mathrm{A} : \mathcal{H} \to \mathcal{K}$ is a bounded linear operator. Define $F = F_1 + F_2 \circ \mathrm{A}$. We assume that:*

A.1 $\mathrm{Im}(\mathrm{A}) \neq \emptyset$.

A.2 $0 \in \mathrm{ri}(\mathrm{dom}(F_1) - \mathrm{A}\mathrm{dom}(F_2))$ *(here finite dimensions).*

A.3 The proximity operator of F_1 and F_2 are simple to compute analytically.

Let $\mu_k \in (0, 2/\|\mathrm{A}\|^2)$. Define the recursion

$$u_{k+1} = \mu_k \left(\mathrm{I} - \mathrm{prox}_{F_2/\mu_k}\right) \left(u_k/\mu_k + \mathrm{A} \circ \mathrm{prox}_{F_1} \left(-\mathrm{A}^* u_k + x\right)\right) \ .$$

Then, $u_k \to u^\star$, and $p_k = \mathrm{prox}_{F_1}(-\mathrm{A}^* u_k + x) \to \mathrm{prox}_F(x)$.

Proximity operator of $F_1 + F_2 \circ \mathrm{A}$

Lemma *Let* $F_1 \in \Gamma_0(\mathcal{H})$ *and* $F_2 \in \Gamma_0(\mathcal{K})$, *and* $\mathrm{A} : \mathcal{H} \to \mathcal{K}$ *is a bounded linear operator. Define* $F = F_1 + F_2 \circ \mathrm{A}$. *We assume that:*

A.1 $\mathrm{Im}(\mathrm{A}) \neq \emptyset$.

A.2 $0 \in \mathrm{ri}(\mathrm{dom}(F_1) - \mathrm{A}\mathrm{dom}(F_2))$ *(here finite dimensions).*

A.3 The proximity operator of F_1 *and* F_2 *are simple to compute analytically.*

Let $\mu_k \in (0, 2/\|\mathrm{A}\|^2)$. *Define the recursion*

$$u_{k+1} = \mu_k \left(\mathrm{I} - \mathrm{prox}_{F_2/\mu_k}\right)\left(u_k/\mu_k + \mathrm{A} \circ \mathrm{prox}_{F_1}\left(-\mathrm{A}^* u_k + x\right)\right) .$$

Then, $u_k \to u^\star$, *and* $p_k = \mathrm{prox}_{F_1}(-\mathrm{A}^* u_k + x) \to \mathrm{prox}_F(x)$.

- Multi-step algorithms can be used as well (on the dual as above).
- The convergence rate can be made precise (linear or $O(1/k^s), s = 1, 2$) under additional assumptions.
- Robustness to errors (see in a little while).
- For $\mathrm{A} = \mathrm{Id}$, other algorithms are possible : Douglas-Rachford or Dykstra algorithm (on the primal).
- Alternative : Augmented Lagrangians and solve by ADMM (for A injective,) or

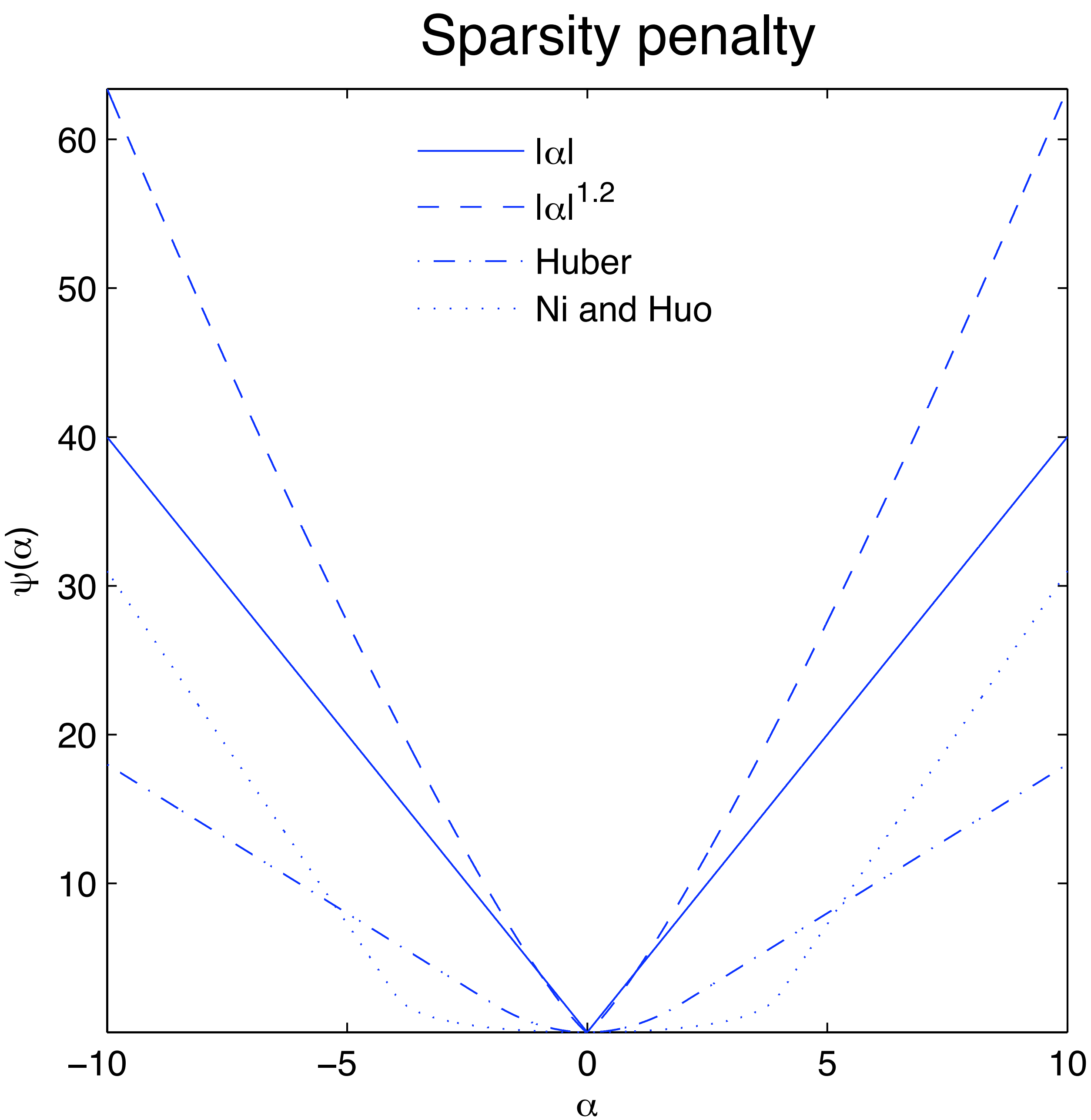
Examples of proximity operators: sparsity penalties

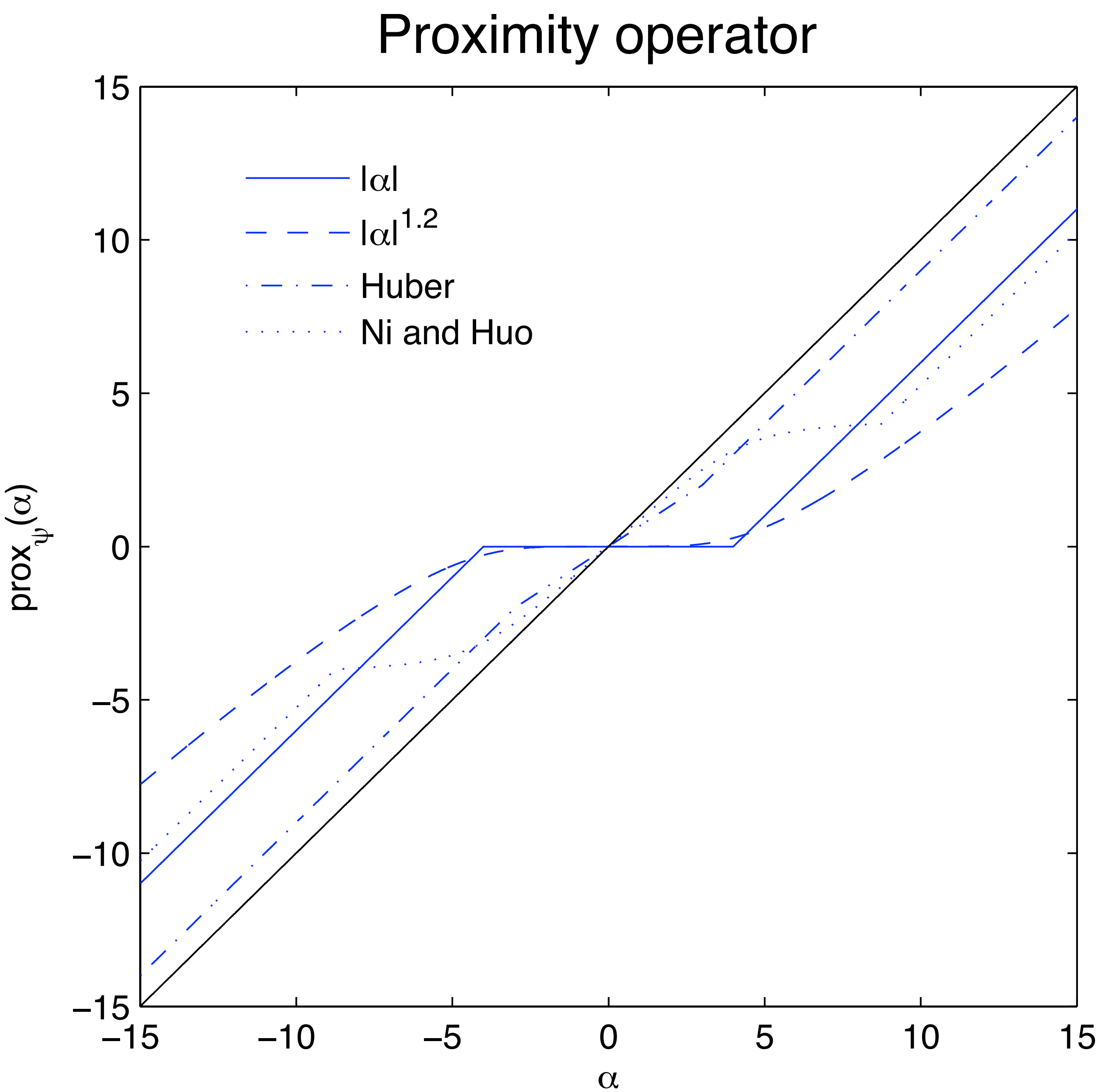
Theorem *Let* $\Psi(x) = \sum_i \psi(x_i)$. *Suppose that* ψ *satisfies, (i)* ψ *is convex even-symmetric, non-negative and non-decreasing on* $[0, +\infty)$, *and* $\psi(0) = 0$. *(ii)* ψ *is twice differentiable on* $\mathbb{R} \setminus \{0\}$. *(iii)* ψ *is continuous on* $\mathbb{R}$, *it is not necessarily smooth at zero and admits a positive right derivative at zero* $\psi'_+(0) = \lim_{h \to 0^+} \frac{\psi(h)}{h} > 0$. *Then, the proximity operator* $\mathrm{prox}_{\kappa\Psi}(x)$ *has exactly one continuous solution decoupled in each coordinate* x_i*:*

$$\hat{x}_i = \mathrm{prox}_{\kappa\psi}(x_i) = \begin{cases} 0 & \text{if } |x_i| \le \kappa\psi'_+(0) \ , \\ x_i - \kappa\psi'(\hat{x}_i) & \text{if } |x_i| > \kappa\psi'_+(0) \ . \end{cases}$$

- Thresholding/shrinkage operator: e.g. soft-thresholding for the ℓ_1 norm.
- Available for many other functions in the literature, either regularization penalties or data fidelity.

Examples


Sparsity penalty
60
50
40
30
20
10
ψ(α)
−10
−5
0
5
10
α
|α|
|α|1.2
Huber
Ni and Huo


Proximity operator
15
10
5
0
−5
−10
−15
proxψ(α)
−15
−10
−5
0
5
10
15
α
|α|
|α|1.2
Huber
Ni and Huo

Outline

- Introduction.
- Non-smooth convex optimization.
  - Elements of convex analysis.
  - Elements of duality.
  - Optimality conditions.
- Proximal framework and operator splitting.
  - Proximal calculus.
  - Monotone operator splitting.
  - Sum of two functions.
  - Generalization to more than two functions.
- Take-away messages.

The gist of splitting

$$(\mathrm{P}): \min_{x \in \mathcal{H}} \sum_{i=1}^{n} F_i(x), \quad F_i : \mathcal{H} \to \mathbb{R} \cup \{+\infty\},\ F_i \in \Gamma_0(\mathcal{H}),\ \text{and } \cap_i \operatorname{dom}(F_i) \neq \emptyset.$$

Theorem

(i) Existence: (P) *possesses at least one solution if* $F = \sum_i F_i$ *is coercive, i.e.* $M^\star \neq \emptyset$.

(ii) Uniqueness: (P) *possesses at most one solution if* F *is strictly convex. This occurs in particular when either one of the* F_i*'s is strictly convex.*

(iii) Characterization: Let $x \in \mathcal{H}$. *Then the following statements are equivalent:*

(a) x *solves* (P).

(b) $x = \operatorname{prox}_{\gamma F}(x)$, $\gamma > 0$, *(proximal algorithm [Martinet 1972]).*

$M^\star \neq \emptyset$ in the sequel to avoid trivialities.

The gist of splitting

$$(\mathrm{P}): \min_{x\in\mathcal{H}} \sum_{i=1}^{n} F_i(x), \quad F_i : \mathcal{H} \to \mathbb{R}\cup\{+\infty\},\ F_i \in \Gamma_0(\mathcal{H}),\ \text{and}\ \cap_i \mathrm{dom}(F_i) \neq \emptyset.$$

Theorem

(i) Existence: (P) *possesses at least one solution if* $F = \sum_i F_i$ *is coercive, i.e.* $M^\star \neq \emptyset$.

(ii) Uniqueness: (P) *possesses at most one solution if* F *is strictly convex. This occurs in particular when either one of the* F_i*'s is strictly convex.*

(iii) Characterization: Let $x \in \mathcal{H}$*. Then the following statements are equivalent:*

(a) x *solves* (P).

(b) $x = \boxed{\mathrm{prox}_{\gamma F}(x)}$, $\gamma > 0$, *(proximal algorithm [Martinet 1972]).*

Explicit computation difficult in general

$M^\star \neq \emptyset$ in the sequel to avoid trivialities.

Monotone operator splitting schemes

- Idea: replace explicit evaluation of $\mathrm{prox}_{\gamma(\sum_i F_i)}$, by a sequence of calculations involving only each $\mathrm{prox}_{\gamma F_i}$ at a time.

$$n = 2$$

Splitting method	Assumptions
Forward-Backward [Gabay 83, Tseng 91]	Either F_1 or F_2 has a Lipschitz-continuous gradient.
Backward-Backward [Lions 78]	F_1, F_2 nonsmooth but do not converge to $(\partial F)^{-1}(0)$, but to $\cap_i (\partial F_i)^{-1}(0)$. Problems with sum of indicator functions or Moreau envelopes.
Douglas/Peaceman-Rachford [Douglas-Rachford 56, Lions-Mercier 79]	F_1, F_2 nonsmooth. Most general.
Alternating-Direction Method of Multipliers (ADMM) [Gabay et al. 80's, Glowinski et al. 70's]	F_1, F_2 nonsmooth, composition by an injective linear operator.
Primal-dual splitting [Arrow-Hurwicz 1956, Chambolle-Pock 2011]	F_1 and F_2 nonsmooth, composition by an arbitrary linear operator.

Monotone operator splitting schemes

- Idea: replace explicit evaluation of $\mathrm{prox}_{\gamma(\sum_i F_i)}$, by a sequence of calculations involving only each $\mathrm{prox}_{\gamma F_i}$ at a time.

$$n > 2$$

Generalized forward-backward [Raguet, Fadili and Peyré, 2013]	F_1 smooth, all others non-smooth.
Spingarn's method (Douglas/Peaceman-Rachford on product spaces), parallel splitting [Spingarn 83, Combettes et al. 08]	All F_i are nonsmooth.
Projective splitting, parallel splitting [Eckstein 09]	All F_i are nonsmooth.
Primal-dual splitting (product pace trick) [Combettes et al. 2011]	All F_i smooth or not, composition by linear operators, infimal-convolution.

Outline

- Introduction.
- Non-smooth convex optimization.
  - Elements of convex analysis.
  - Elements of duality.
  - Optimality conditions.
- Proximal framework and operator splitting.
  - Proximal calculus.
  - Monotone operator splitting.
  - Sum of two functions.
  - Generalization to more than two functions.
- Take-away messages.

Forward-Backward: the gist

$$(\mathsf{P}): \min_{x \in \mathcal{H}} F_1(x) + F_2(x),$$

- $F_i : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$, $F_i \in \Gamma_0(\mathcal{H})$;
- $\cap_i \operatorname{dom}(F_i) \neq \emptyset$;
- Set of minimizers $M^\star$ is nonempty (e.g. by coercivity) ;
- F_2 has a β-Lipschitz gradient.

$$\begin{aligned}
& x \text{ is a (global) minimizer of} (\mathsf{P}) \\
\iff \quad & 0 \in \partial(F_1 + F_2)(x) \\
\iff \quad & -\nabla F_2(x) \in \partial F_1(x) \\
\iff \quad & (x - \mu \nabla F_2(x)) - x \in \partial(\mu F)(x) \\
\iff \quad & x = \underbrace{\operatorname{prox}_{\mu F_1}}_{\text{Backward step}} \underbrace{(x - \mu \nabla F_2(x))}_{\text{Forward step}} \\
\iff \quad & x \in \operatorname{Fix}\left(\operatorname{prox}_{\mu F_1} \circ (\mathrm{I} - \mu \nabla F_2)\right) .
\end{aligned}$$

Forward-Backward: the scheme

Initialization : choose some $x_0 \in \mathrm{dom}(F)$, a sequence or a fixed $\mu_k \in (0, 2/\beta)$.

Main iteration :

repeat

1. *Gradient descent (forward) step :*

$$x_{k+1/2} = x_k - \mu_k \nabla F_2(x_k).$$

2. *Proximal (backward) step :*

$$x_{k+1} = \mathrm{prox}_{\mu_k F_1}\left(x_{k+1/2}\right) .$$

$k \leftarrow k+1.$

until *convergence*;

Forward-Backward: convergence

Theorem *Suppose that F_1 and $F_2 \in \Gamma_0(\mathcal{H})$, and F_2 has a β-Lipschitz continuous gradient. Let $(\mu_k)_{k\in\mathbb{N}}$ be a sequence such that $0 < \inf_k \mu_k \leq \sup_k \mu_k < 2/\beta$, let $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$ be error sequences in $\mathcal{H}$ such that $\sum_k \|a_k\| < +\infty$ and $\sum_k \|b_k\| < +\infty$. Fix $x_0 \in \mathcal{H}$, and define the sequence of iterates :*

$$x_{k+1} = (1-\lambda_k)x_k + \lambda_k \left(\operatorname{prox}_{\mu_k F_1}\left(x_k - \mu_k\left(\nabla F_2(x_k)\right) + b_k\right)\right) + a_k\Big)$$

where $\lambda_k \in]0,1]$. Then, $(x_k)_{k\in\mathbb{N}}$ converges to a minimizer of (P).

Forward-Backward: convergence

Theorem *Suppose that F_1 and $F_2 \in \Gamma_0(\mathcal{H})$, and F_2 has a β-Lipschitz continuous gradient. Let $(\mu_k)_{k\in\mathbb{N}}$ be a sequence such that $0 < \inf_k \ \mu_k \le \sup_k \ \mu_k < 2/\beta$, let $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$ be error sequences in $\mathcal{H}$ such that $\sum_k \|a_k\| < +\infty$ and $\sum_k \|b_k\| < +\infty$. Fix $x_0 \in \mathcal{H}$, and define the sequence of iterates :*

$$x_{k+1} = (1-\lambda_k)x_k + \lambda_k \left(\mathrm{prox}_{\mu_k F_1}\left(x_k - \mu_k\left(\nabla F_2(x_k)\right) + b_k\right)\right) + a_k\right)$$

where $\lambda_k \in]0,1]$. Then, $(x_k)_{k\in\mathbb{N}}$ converges to a minimizer of (P).

Theorem *Consider the errorless and unrelaxed version of the above forward-backward algorithm. Then, the objective converges at the rate $1/k$. If F is strongly convex, then the convergence is linear on the iterate and objective.*

Forward-Backward: convergence

Theorem *Suppose that F_1 and $F_2 \in \Gamma_0(\mathcal{H})$, and F_2 has a β-Lipschitz continuous gradient. Let $(\mu_k)_{k\in\mathbb{N}}$ be a sequence such that $0 < \inf_k \ \mu_k \leq \sup_k \ \mu_k < 2/\beta$, let $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$ be error sequences in $\mathcal{H}$ such that $\sum_k \|a_k\| < +\infty$ and $\sum_k \|b_k\| < +\infty$. Fix $x_0 \in \mathcal{H}$, and define the sequence of iterates :*

$$x_{k+1} = (1-\lambda_k)x_k + \lambda_k \left(\mathrm{prox}_{\mu_k F_1}\left(x_k - \mu_k\left(\nabla F_2(x_k)\right) + b_k\right)\right) + a_k\right)$$

where $\lambda_k \in]0,1]$. Then, $(x_k)_{k\in\mathbb{N}}$ converges to a minimizer of (P).

Theorem *Consider the errorless and unrelaxed version of the above forward-backward algorithm. Then, the objective converges at the rate $1/k$. If F is strongly convex, then the convergence is linear on the iterate and objective.*

- Robustness to errors in the proximity operator and in the gradient.
- $1/k$ convergence rate in the objective : nothing surprising as a one-memory first-order scheme (recall projected gradient descent).
- Can we attain the complexity upper-bound rate $1/k^2$? Yes : multistep scheme by [Nesterov 2007,Beck-Teboulle 09,Tseng 09,Chambolle-Dossal 16].

FISTA scheme

Initialization : choose some $x_0 \in \mathrm{dom}(F)$, a sequence or a fixed $\mu_k \in]0, 1/\beta]$, $k = 1$, $a \geq 2$.

Main iteration :

repeat

$$y_k = x_k + \frac{k-1}{k+a}(x_k - x_{k-1}).$$

$$x_{k+1} = \mathrm{prox}_{\mu_k F_1}\left(y_k - \mu_k \nabla F_2(y_k)\right) .$$

$k \leftarrow k + 1$

until *convergence*;

FISTA scheme: Convergence

Theorem *Consider the FISTA algorithm with the same assumptions as before.*

1. *If $a = 2$: $F(x_k) - F(x^\star) = O(1/k^2)$. If F is strongly convex, then the convergence is linear with a better rate than the forward-backward.*
2. *If $a > 2$: then*
   (a) *x_k converges to a minimizer of (P).*
   (b) *$F(x_k) - F(x^\star) = o(1/k^2)$.*

- Robustness to errors but may degrade the rates.
- $1/k^2$ in the objective is optimal for first-order schemes on this class of problems.

Douglas-Rachford: the gist

$$(\mathsf{P}): \min_{x \in \mathcal{H}} F_1(x) + F_2(x),$$

- $F_i : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$, $F_i \in \Gamma_0(\mathcal{H})$;
- $\cap_i \mathrm{ri}(\mathrm{dom}(F_i)) \neq \emptyset$;
- Set of minimizers $M^\star$ is nonempty (e.g. by coercivity) ;

$$
\begin{aligned}
& x \text{ is a (global) minimizer of} (\mathsf{P}) \\
\iff \quad & 0 \in \partial(F_1 + F_2)(x) \\
\iff \quad & \exists z \in \mathcal{H}, z - x \in \partial(\gamma F_1(x) \text{ and } x - z \in \partial(\gamma F_2)(x) \ , \quad \gamma > 0 \\
\iff \quad & x = \mathrm{prox}_{\gamma F_1}(z) \text{ and } (2x - z) - x \in \partial(\gamma F_2)(x) \\
\iff \quad & x = \mathrm{prox}_{\gamma F_1}(z) \text{ and } x = \mathrm{prox}_{\gamma F_2}(2x - z) = \mathrm{prox}_{\gamma F_2} \circ \mathrm{rprox}_{\gamma F_1}(z) \\
\iff \quad & x = \mathrm{prox}_{\gamma F_1}(z) \text{ and } z = 2\,\mathrm{prox}_{\gamma F_2} \circ \mathrm{rprox}_{\gamma F_1}(z) - (2x + z) \\
\iff \quad & x = \mathrm{prox}_{\gamma F_1}(z) \text{ and } z = 2\,\mathrm{prox}_{\gamma F_2} \circ \mathrm{rprox}_{\gamma F_1}(z) - \mathrm{rprox}_{\gamma F_1}(z) \\
\iff \quad & x = \mathrm{prox}_{\gamma F_1}(z) \text{ and } z = \left(1 - \frac{\lambda}{2}\right) z + \frac{\lambda}{2} \mathrm{rprox}_{\gamma F_2} \circ \mathrm{rprox}_{\gamma F_1}(z) \ , \quad \lambda \in [0, 2] \\
\iff \quad & z \in \mathrm{Fix}\left(\left(1 - \frac{\lambda}{2}\right) \mathrm{I} + \frac{\lambda}{2} \mathrm{rprox}_{\gamma F_2} \circ \mathrm{rprox}_{\gamma F_1}\right) \text{ and } x = \mathrm{prox}_{\gamma F_1}(z) \in M^\star .
\end{aligned}
$$

Douglas-Rachford: the scheme

Initialization : choose some $x_0 \in \mathcal{H}$, $\lambda_k \in (0,2)$, $\gamma > 0$.

Main iteration :

repeat

1. *First proximity operator :* Compute

$$z_{k+1/2} = 2\,\mathrm{prox}_{\gamma F_1}(z_k) - z_k\ .$$

2. *Second proximity operator :*

$$z_{k+1} = (1-\lambda_k/2)z_k + \lambda_k/2\left(2\,\mathrm{prox}_{\gamma F_2}\left(z_{k+1/2}\right) - z_{k+1/2}\right)\ .$$

$k \leftarrow k+1$.

until *convergence*;

Douglas-Rachford: Convergence

Theorem *Let* $\gamma \in (0, +\infty)$, *let* $(\lambda_k)_{k\in\mathbb{N}}$ *be a sequence in* $(0, 2)$, *and let* $(a_k)_{k\in\mathbb{N}}$ *and* $(b_k)_{k\in\mathbb{N}}$ *be sequences in* $\mathcal{H}$ *such that* $\sum_{k\in\mathbb{N}} \lambda_k(2-\lambda_k) = +\infty$ *and* $\sum_{k\in\mathbb{N}} \lambda_k \left(\|a_k\| + \|b_k\|\right) < +\infty$. *Fix* $x_0 \in \mathrm{dom}(F)$ *and define the sequence of iterates,*

$$\begin{aligned} z_{k+1/2} &= \mathrm{prox}_{\gamma F_1}(z_k) + b_k \ , \\ z_{k+1} &= z_k + \lambda_k \left(\mathrm{prox}_{\gamma F_2} \circ \left(2z_{k+1/2} - z_k\right) + a_k - z_{k+1/2}\right) \ . \end{aligned}$$

Then z_k *converges to some fixed point* $z^\star$ *and* $x^\star = \mathrm{prox}_{\gamma F_1}(z^\star) \in M^\star$.

- Again, robustness to errors in both proximity operators.
- Convergence rates in a variety of situations : asymptotic regularity, under strong convexity, partial smoothness [Liang, Fadili and Peyré 2015,Liang, Fadili and Peyré 2015, 2017].

ADMM (DR on the dual): the gist

$$(\mathsf{P}): \inf_{x\in\mathcal{H}} F(x) + G\circ \mathrm{A}(x) \iff (\mathsf{P}^*): \min_{u\in\mathcal{K}} F^*\circ(-\mathrm{A}^*)(u) + G^*(u),$$

- $F \in \Gamma_0(\mathcal{H})$, $G \in \Gamma_0(\mathcal{K})$;
- $\mathrm{A} : \mathcal{H} \to \mathcal{K}$ bounded and injective linear operator ;
- Domain qualification condition ;
- $M^\star \neq \emptyset$.

Remember composition lemma

ADMM (DR on the dual): the gist

$$(\mathsf{P}) : \inf_{x \in \mathcal{H}} F(x) + G \circ \mathrm{A}(x) \iff (\mathsf{P}^*) : \min_{u \in \mathcal{K}} F^* \circ (-\mathrm{A}^*)(u) + G^*(u),$$

- $F \in \Gamma_0(\mathcal{H})$, $G \in \Gamma_0(\mathcal{K})$;
- $\mathrm{A} : \mathcal{H} \to \mathcal{K}$ bounded and injective linear operator ;
- Domain qualification condition ;
- $M^\star \neq \emptyset$.

- Solve (P) : Apply DR to (P^*).
  - Use Fenchel-Rockafellar duality to compute the proximity operator of $F^* \circ (-A^*)$: injectivity important to ensure strong monotonicity hence uniqueness of the minimizer in x.

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{H}} F(x) + \langle u_k, \mathrm{A}x \rangle + \frac{\gamma}{2} \|\mathrm{A}x - v_k\|^2 ,$$

Remember composition lemma

  - Use Fenchel-Rockafellar duality to compute the proximity operator of G^* (in fact Moreau identity).

$$v_{k+1} = \operatorname*{argmin}_{v \in \mathcal{K}} G(v) - \langle u_k, v \rangle + \frac{\gamma}{2} \|\mathrm{A}x_{k+1} - v\|^2 = \operatorname{prox}_{G/\gamma} (\mathrm{A}x_{k+1} + u_k/\gamma) ,$$

  - Update dual variable.

$$u_{k+1} = u_k + \gamma (\mathrm{A}x_{k+1} - v_{k+1}) .$$

- Minimizes the augmented Lagrangian function associated to (P).

ADMM: Convergence

Theorem *Let the convex program* (P)*, where* A *is injective. Let* $\gamma \in (0, +\infty)$*, and* $(a_k)_{k\in\mathbb{N}}$ *and* $(b_k)_{k\in\mathbb{N}}$ *be summable sequences in* $\mathcal{H}$ *and* $\mathcal{K}$*. Solve* (P) *using the ADMM, where the sub-problems for updating* x_k *and* v_k *are solved either exactly or with errors at most* a_k *and* b_k*. Then if* (P) *has a Kuhn-Tucker pair,* x_k *converges to a solution of* (P) *and* u_k *converges to a solution of the dual problem* (P^*)*.*

- Again, robustness to errors in both proximity operators.
- Convergence rates in a variety of situations : asymptotic regularity, under strong convexity, partial smoothness [Liang, Fadili and Peyré 2015,Liang, Fadili and Peyré 2015, 2017].
- Flexibility in the choice of splitting to ensure injectivity.

Primal-dual splitting: the gist

$$(\mathsf{P}): \inf_{x\in\mathcal{H}} F(x) + G\circ \mathrm{A}(x) \iff (\mathsf{P}^*): \min_{u\in\mathcal{K}} F^*\circ(-\mathrm{A}^*)(u) + G^*(u),$$

- $F \in \Gamma_0(\mathcal{H})$, $G \in \Gamma_0(\mathcal{K})$;
- $\mathrm{A} : \mathcal{H} \to \mathcal{K}$ a linear operator ;
- Domain qualification condition ;
- $M^\star \neq \emptyset$.

Lemma (x, u) *is a Kuhn-Tucker pair if and only if*

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \underbrace{\begin{pmatrix} \partial F & 0 \\ 0 & \partial G^* \end{pmatrix}}_{T_1} \begin{pmatrix} x \\ u \end{pmatrix} + \underbrace{\begin{pmatrix} 0 & \mathrm{A}^* \\ -\mathrm{A} & 0 \end{pmatrix}}_{T_2} \begin{pmatrix} x \\ u \end{pmatrix} .$$

T_1 and T_2 are maximal monotone, and T_2 is skew-adjoint linear.

- T_2 is Lipschitz but not co-coercive $\Rightarrow$ forward-backward does **not** apply.
- Compensate for lack of co-coercivity :
  - Forward-Backward-Forward [Tseng 98].
  - Forward-backward in a different metric [Chambolle-Pock 2011, Yuan-He 2011].

Primal-dual splitting: the gist

$$(\mathsf{P}): \inf_{x\in\mathcal{H}} F(x) + G \circ \mathrm{A}(x) \iff (\mathsf{P}^*): \min_{u\in\mathcal{K}} F^* \circ (-\mathrm{A}^*)(u) + G^*(u),$$

- $F \in \Gamma_0(\mathcal{H})$, $G \in \Gamma_0(\mathcal{K})$;
- $\mathrm{A} : \mathcal{H} \to \mathcal{K}$ a linear operator ;
- Domain qualification condition ;
- $M^\star \neq \emptyset$.

- A preconditioned version of ADMM [Chambolle-Pock 2011].
- The trick is to precondition the update of x_{k+1}, $\tau\gamma < 1/|||\mathrm{A}|||^2$:

$$x_{k+1} = \operatorname*{argmin}_{x\in\mathcal{H}} F(x) + \langle u_k, \mathrm{A}x\rangle + \frac{\gamma}{2}\|\mathrm{A}x - v_k\|^2 + \frac{1}{2}\Big\langle\Big(\frac{1}{\tau} - \gamma\mathrm{A}\mathrm{A}^*\Big)(x - x_k), x - x_k\Big\rangle,$$

- This is equivalent to :

$$x_{k+1} = \operatorname{prox}_{\tau F}(x_k - \tau\mathrm{A}^*\bar{x}_k), \bar{x}_k := u_k + \gamma(\mathrm{A}x_k - v_k).$$

- Other steps remain unchanged.

Primal-Dual splitting: Convergence

Theorem *Consider the convex program* (P) *where* A *is a bounded linear operator. Let* $\gamma \in (0, +\infty)$ *and* $\tau\sigma < 1/|||\mathrm{A}|||^2$*. Assume that* (P) *has a Kuhn-Tucker point and solve it with the pre-conditioned ADMM. Then the sequence of primal and dual pair converges to Kuhn-Tucker point. Furthermore, the (partial) restricted gap converges at the rate* $O(1/k)$*.*

- Applicable algorithm to a wide spectrum of problems.
- Robustness to erros as well.
- Can be accelerated with multi-step schemes for strongly convex objectives.

Outline

- Introduction.
- Non-smooth convex optimization.
  - Elements of convex analysis.
  - Elements of duality.
  - Optimality conditions.
- Proximal framework and operator splitting.
  - Proximal calculus.
  - Monotone operator splitting.
  - Sum of two functions.
  - Generalization to more than two functions.
- Take-away messages.

Spingarn's method: the gist

$$(\mathsf{P}) : \min_{x \in \mathcal{H}} \sum_{i=1}^{n} F_i(x),$$

- $F_i : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$, $F_i \in \Gamma_0(\mathcal{H})$;
- $\cap_i \mathrm{ri}(\mathrm{dom}(F_i)) \neq \emptyset$;
- Set of minimizers $M^\star$ is nonempty (e.g. by coercivity) ;

Spingarn's method: the gist

$$(\mathrm{P}) : \min_{x \in \mathcal{H}} \sum_{i=1}^{n} F_i(x),$$

- Define the closed subspace $\mathcal{S} = \{(x_1, \cdots, x_n) \in \mathcal{H}^n : \sum_i x_i = 0\}$, and its orthogonal complement $\mathcal{S}^\perp = \{(x_1, \cdots, x_n) \in \mathcal{H}^n : x_1 = x_2 = \cdots = x_n\}$.
- Let $\mathcal{N}_{\mathcal{S}^\perp}$ be its normal cone, i.e. subdifferential of $\imath_{\mathcal{S}^\perp}$.
- (P) is equivalent to $\min_{(x_1,\cdots,x_n)} \sum_{i=1}^n F_i(x_i) + \imath_{\mathcal{S}^\perp}(x_1, \cdots, x_n)$.
- Let's remark that $\partial(\sum_i F_i(x_i)) = \partial F_1(x_1) \times \cdots \times \partial F_n(x_n)$. Thus

$$\begin{aligned}
& 0 \in F(x) \\
\iff \quad & 0 \in \times_i \partial F_i(x_i) + \mathcal{N}_{\mathcal{S}}(x_1, \cdots, x_n) \\
\iff \quad & x_1 = \cdots = x_n, \exists u_i = \partial F_i(x_i), \sum_i u_i = 0 \, .
\end{aligned}$$

- Applying the Douglas-Rachford splitting to this problem produces Spingarn's method :
  - perform independent proximal steps on each of the functions F_i (separable, and so are the proximity operators) ;
  - and then compute the next iterate by essentially averaging the results.

Douglas-Rachford for $n>2$: the scheme

Initialization : Choose $(y_0^i)_{1\leq i\leq n} \in \mathcal{H}^n$, $\gamma \in (0,+\infty)$, weights $w_i \in (0,1]$ that sum up to 1 (e.g. $1/n$), and let $x_0 = \sum_{i=1}^n w_i y_{0i}$.

Main iteration :

repeat

1. Compute the proximal operators (in parallel if desired) :

   for $i = 1$ *to* n **do**

   $$z_k^i = \operatorname{prox}_{\gamma w_i F_i} y_k^i \ .$$

2. Average the results :

   $$x_{k+1} = \sum_{i=1}^n w_i z_k^i \ .$$

3. Second proximal step of Douglas-Rachford :

   for $i = 1$ *to* n **do**

   $$y_{k+1}^i = y_k^i + 2x_{k+1} - x_k - z_k^i \ .$$

until *convergence*;

Douglas-Rachford for $n>2$: Convergence

Theorem *Let $\gamma \in (0, +\infty)$, let $\left(a_k^i\right)_{ink\in\mathbb{N}}$ be the sequence of errors in each proximity operator $\operatorname{prox}_{\gamma F_i}(x_k)$ such that $\sum_{k\in\mathbb{N}} \left\|a_k^i\right\| < +\infty$ for each $i = 1, \cdots, n$. If the functions F_i satisfy a qualification condition on the intersection of the relative interior of their domains, then x_k converges to $x^\star$, a solution of (P).*

Douglas-Rachford for $n>2$: Convergence

Theorem *Let* $\gamma \in (0, +\infty)$*, let* $\left(a_k^i\right)_{ink\in\mathbb{N}}$ *be the sequence of errors in each proximity operator* $\operatorname{prox}_{\gamma F_i}(x_k)$ *such that* $\sum_{k\in\mathbb{N}} \left\|a_k^i\right\| < +\infty$ *for each* $i = 1, \cdots, n$*. If the functions* F_i *satisfy a qualification condition on the intersection of the relative interior of their domains, then* x_k *converges to* $x^\star$*, a solution of* (P)*.*

- Again, robustness to errors in each proximity operator.
- Convergence rates in [Liang, Fadili and Peyré 2015].

Many, many structured optimization problems can be solved within this framework:
see practical work sessions

Take away messages

- Convex analysis and proximal splitting are a powerful framework for solving convex optimization problems, non-necessarily smooth.
- Good and fast solvers for large-scale problems with grounded theoretical results.
- A wide variety of applications.
- Try it to be convinced.

Bibliography

•C. Lemaréchal and J.-B. Hiriart-Urruty, Convex Analysis and Minimization Algorithms I and II, Springer, 2nd ed., 1996.

•D. Gabay, Applications of the method of multipliers to variational inequalities, in Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems, M. Fortin and R. Glowinski, eds., North-Holland, Amsterdam, 1983.

•R.T. Rockafellar, Convex analysis, Princeton University Press, 1970.

•R.T. Rockafellar, The Theory of Subgradients and its Applications to Problems of Optimization: Convex and Nonconvex Functions, Helderman-Verlag, Berlin, 1981.

•R. T. Rockafellar and J. B. Wets, Variational analysis, Springer-Verlag, New York, 1997.

•I. Ekeland and R. Temam, Analyse convexe et problèmes variationnels, Dunod, 1974.

•R. Glowinski, J. Lions, and R. Trémolières, Analyse numérique des inéquations variationnelles, vol. 1, Dunod, Paris, 1 ed., 1976.

•H.H. Bauschke and P.L. Combettes: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer-Verlag, 2011.

•H.H. Bauschke, R.S. Burachik, P.L. Combettes, V. Elser, D.R. Luke, and H. Wolkowicz (editors): Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer-Verlag, 2011.

•J-L. Starck, F. Murtagh, M.J. Fadili, "Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity" , Cambridge University Press, 2010.

Thanks
Any questions ?