Lecture 2: Convex non-smooth optimisation

Luca Calatroni
CR CNRS, Laboratoire I3S
CNRS, UCA, Inria SAM, France

MIVA ERASMUS BIP PhD winter school

Advanced methods for mathematical image analysis

University of Bologna, IT

January 18-20 2022

Table of contents

Life is not smooth...

In many applications the function g in

$$\min_{x \in \mathbb{R}^n} \; \{F(x) := f(x) + g(x)\},$$

is different from 0. Typically, g is convex, but **non differentiable** so its gradient (and henceforth the one of F) cannot be defined in a standard way.

Note: take **implicit gradient-descent** for suitable $\tau > 0$:

$$x_{k+1} = x_k - \tau \nabla f(x_{k+1}) \quad \Leftrightarrow \quad \nabla f(x_{k+1}) + \frac{x_{k+1} - x_k}{\tau} = 0,$$

So if x_{k+1} exists, it is a critical point of the function:

$$x \mapsto f(x) + \frac{\|x_k - x\|^2}{2\tau}$$

If $f \in \Gamma_0(\mathbb{R}^n)$ (not necessarily smooth!), x_{k+1} is indeed the **unique critical point** of this function...

non-smoothness encoded via "implicit" updates?

Non-smooth optimisation

Non-smooth optimisation

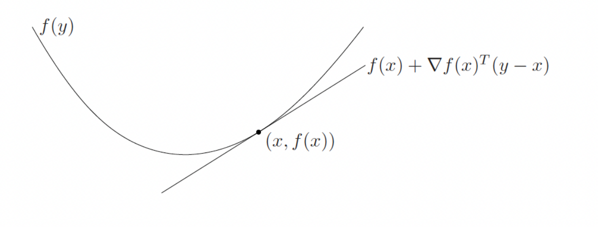
Subgradients

A preliminary observation

One can show that if $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is differentiable:

$$f \text{ is convex} \quad \Leftrightarrow \quad (\forall x, y \in \mathbb{R}^n) \quad f(y) \geq \underbrace{f(x) + \nabla f(x)^T (y - x)}_{=:\phi(y;x)}$$

- the function $\phi(\cdot; x)$ is an affine lower bound/estimator of $f(\cdot)$
- the tangent to f at any $x \in \text{dom}(f)$ is below f at all points.

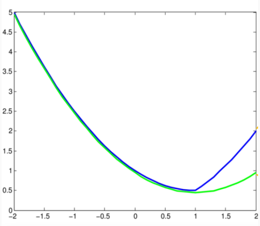


A preliminary observation

One can show that if $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is differentiable:

$$f \text{ is convex} \quad \Leftrightarrow \quad (\forall x, y \in \mathbb{R}^n) \quad f(y) \geq \underbrace{f(x) + \nabla f(x)^T (y - x)}_{=:\phi(y;x)}$$

- the function $\phi(\cdot; x)$ is an affine lower bound/estimator of $f(\cdot)$
- the tangent to f at any $x \in \text{dom}(f)$ is below f at all points.



Recall: If f is μ-strongly convex, then, analogously, f has a quadratic lower bound

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n.$$

Subgradients and subdifferential

Definition (Subgradients and subdifferential)

Let $g \in \mathcal{P}$ be **convex**. Then, a vector $p \in \mathbb{R}^n$ is a *subgradient* of g at point $x \in \text{dom}(g)$ iff:

$$g(y) \geq g(x) + p^T(y - x), \qquad \forall y \in \mathbb{R}^n$$

If $x \notin \text{dom}(g)$, we set $\partial g(x) = \emptyset$. The set of all subgradients at a point $x \in \mathbb{R}^n$ is called the *subdifferential* of g in x, and it is the denoted by:

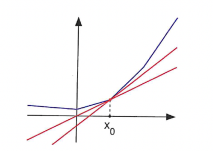
$$\partial g(x) = \{p \in \mathbb{R}^n : p \text{ is a subgradient of } g \text{ at point } x\}$$

Interpretation:

- $p \in \partial g(x)$ if and only if $\phi(y; x) = g(x) + p^T(y - x)$ is a lower affine bound for g.
- $\partial g(x)$ collects all the **slopes** of the tangent lines through x.

Remarks

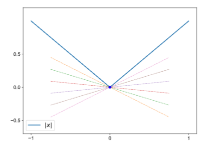
In general, $\partial g(\cdot) : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ is not a singleton

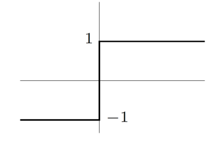


Multiple subgradients at a non-differentiable point x_0.

Example: $g : \mathbb{R} \to \overline{\mathbb{R}}$, $g(x) = |x|$.

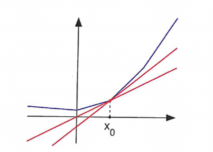
$$\partial g(x) = \begin{cases} \{1\} & \text{if} \quad x > 0 \\ \{-1\} & \text{if} \quad x < 0 \\ [-1, 1] & \text{if} \quad x = 0. \end{cases}$$





Remarks

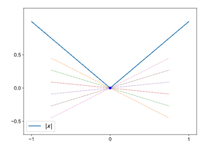
In general, $\partial g(\cdot) : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ is not a singleton

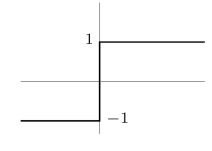


Multiple subgradients at a non-differentiable point x_0.

Example: $g : \mathbb{R} \to \overline{\mathbb{R}}$, $g(x) = |x|$.

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





Proposition (subdifferential at differentiable points)

If g is convex and differentiable in $x \in \mathrm{dom}(g)$, then:

$$\partial g(x) = \{\nabla g(x)\}\,.$$

Subdifferential of norm

Compute $\partial\|x\|$ for all $x \in \mathbb{R}^n$.

- $g(x) = \|x\|$ is differentiable for all $x \neq 0$. There, $\partial\|x\| = \frac{x}{\|x\|}$.
- The point of interest (non-differentiability) is 0

In $x = 0$ subgradients $p \in \mathbb{R}^n$ verify:

$$\|y\| \geq 0 + p^T(y - 0) = p^T y \quad \forall y \in \mathbb{R}^n$$

Take the maximum on both sides for all $y : \|y\| \leq 1$, you get:

$$1 = \max_{y:\|y\|\leq 1} \|y\| \geq \max_{y:\|y\|\leq 1} p^T y = \|p\|$$

Contrarily, if $\|p\| \leq 1$, then by Cauchy-Schwarz inequality there holds:

$$p^T y \leq \|p\|\|y\| \leq \|y\|$$

Hence, we proved $p \in \partial\|0\|$ if and only if $\|p\| \leq 1$. Hence

$$\partial\|0\| = \{p \in \mathbb{R}^n : \|p\| \leq 1\} = B_1(0) \quad \Rightarrow \quad \partial\|x\| = \begin{cases} \frac{x}{\|x\|} & x \neq 0 \\ B_1(0) & x = 0. \end{cases}$$

Calculus rules: separable functions

Often, the n-dimensional function you deal with, can be nicely expressed as the sum of 1D components. For instance, think of:

- **norms** $\|x\|_p^p, p \geq 1$: $\|x\|_p^p = \sum_{i=1}^n |x_i|^p \ldots$
- **sum of norms**, e.g. $g(x) = \|x\|_1 + \frac{\lambda}{2}\|x\|_2^2 = \sum_{i=1}^n \left(|x_i| + \lambda |x_i|^2\right)$.
- ...

Calculus rules: separable functions

Often, the n-dimensional function you deal with, can be nicely expressed as the sum of 1D components. For instance, think of:

- **norms** $\|x\|_p^p, p \geq 1$: $\|x\|_p^p = \sum_{i=1}^n |x_i|^p \dots$
- **sum of norms**, e.g. $g(x) = \|x\|_1 + \frac{\lambda}{2}\|x\|_2^2 = \sum_{i=1}^n \left(|x_i| + \lambda |x_i|^2\right)$.
- ...

Definition (separable function)

Let $g \in \mathcal{P}$ be convex. We say that g is *separable* if there exist proper, univariate convex functions $g_i : \mathbb{R} \to \overline{\mathbb{R}}$ such that

$$g(x) = \sum_{i=1}^n g_i(x_i), \qquad \forall x \in \mathbb{R}^n.$$

Proposition (subdifferential of separable functions)

Let $g \in \mathcal{P}$ be convex and separable. Then, for all $x \in \mathrm{dom}(g)$:

$$\partial g(x) = (\partial g_i(x_i))_{i=1}^n = (\partial g_1(x_1)) \times \dots \times (\partial g_n(x_n)).$$

Calculus rules: sum and multiplication by scalar

Proposition (Moreau-Rockafellar)

Let $g, g_2 : \mathbb{R}^n \to \overline{\mathbb{R}}$ be two proper convex functions. Then:

$$\partial g_1(x) + \partial g_2(x) \subset \partial\left(g_1(\cdot) + g_2(\cdot)\right)(x).$$

Moreover, if $\text{int}(dom(g_1)) \cap \text{int}(dom(g_2)) \neq \emptyset$, then for all $x \in \mathbb{R}^n$:

$$\partial g_1(x) + \partial g_2(x) = \partial\left(g_1(\cdot) + g_2(\cdot)\right)(x).$$

For $\lambda \in \mathbb{R}_{++}$, there holds:

$$\partial\left(\lambda f\right)(x) = \lambda \partial f(x), \qquad \forall x \in \mathbb{R}^n.$$

Example: $\partial(g_1(\cdot) + g_2(\cdot))(x)$ may differ indeed from $\partial g_1(x) + \partial g_2(x)$! In $\mathbb{R}$ take:

$$g_1(x) := \begin{cases} 0 & \text{if } x \leq 0 \\ +\infty & \text{if } x > 0. \end{cases} \qquad g_2(x) := \begin{cases} +\infty & \text{if } x < 0 \\ -\sqrt{x} & \text{if } x \geq 0. \end{cases}$$

We have:

$$\partial g_1(x) = \begin{cases} 0 & \text{if } x < 0 \\ [0, +\infty) & \text{if } x = 0 \\ \emptyset & \text{if } x > 0 \end{cases} \qquad \partial g_2(x) = \begin{cases} \emptyset & \text{if } x \leq 0 \\ -\frac{1}{2\sqrt{x}} & \text{if } x > 0. \end{cases}$$

Hence, $\partial g_1(x) + \partial g_2(x) = \emptyset$ for all $x \in \mathbb{R}$. However, $g_1(x) + g_2(x) = \iota_0(x)$ and $\partial \iota_0(0) = \mathbb{R}$.

Composite subgradients and chain rule

Proposition

Let $f \in \Gamma_0(\mathbb{R}^n)$ be differentiable at $x \in \mathbb{R}^n$ and let $g \in \Gamma_0(\mathbb{R}^n)$, then:

$$\partial(f+g)(x) = \{\nabla f(x)\} + \partial g(x).$$

Proposition

Let $L \in \mathbb{R}^{N \times n}$ and $g : \mathbb{R}^N \to \bar{\mathbb{R}}$ a proper convex function. Then:

$$(\forall x \in \mathbb{R}^n) \quad L^T \partial g(Lx) \subset \partial(g \circ L)(x).$$

Moreover, if $\text{int}(dom(g) \cap R(L) \neq \emptyset$, then:

$$(\forall x \in \mathbb{R}^n) \quad L^T \partial g(Lx) = \partial(g \circ L)(x).$$

Optimality conditions

Analogous to Fermat's rule in non-smooth case.

Theorem (optimality conditions in non-smooth, convex case)

Let $g \in \Gamma_0(\mathbb{R}^n)$. Then:

$$x^* \in \underset{x \in \mathbb{R}^n}{\arg\min}\, g(x) \qquad \Longleftrightarrow \qquad 0 \in \partial g(x^*).$$

Interpretation:

- If the vector $0 \in \mathbb{R}^n$ belongs to $\partial g(x^*)$ ("flat plot"), then x^* is a minimiser.
- If g is differentiable, the result reads $0 = \nabla g(x^*)$ (Fermat's rule).

Stationary points

If $f, g \in \Gamma_0(\mathbb{R}^n)$ and f is smooth

$$\arg\min_{x \in \mathbb{R}^n} \; \{F(x) := f(x) + g(x)\}$$

$$x^* \in \arg\min_{x \in \mathbb{R}^n} F(x) \Leftrightarrow 0 \in \partial F(x^*) = \underbrace{\partial f(x^*)}_{f \text{ is smooth}} + \partial g(x^*) = \{\nabla f(x^*)\} + \partial g(x^*)$$

Definition (stationary point)

A point $x^* \in \mathbb{R}^n$ verifying:

$$0 \in \{\nabla f(x^*)\} + \partial g(x^*) \quad \Leftrightarrow \quad -\nabla f(x^*) \in \partial g(x^*)$$

is said to be a **stationary point** of the composite functional $F := f + g$.

Non-smooth optimisation

The proximal operator

The proximal operator: definition

Crucial tool for the development of non-smooth optimisation algorithms. Relations with activation functions in the context of deep networks (Combettes, Pesquet, '20).

Definition

Let $g \in \mathcal{P}$. Then, the *proximal operator* of g with parameter $\gamma > 0$ is defined as the **multi-valued map** $\text{prox}_{\gamma g} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ defined for all $x \in \mathbb{R}^n$:

$$\text{prox}_{\gamma g}(x) := \arg\min_{y \in \mathbb{R}^n} \underbrace{g(y) + \frac{1}{2\gamma}\|y - x\|^2}_{=: h(y;x)}$$

With no further conditions on g, $\text{prox}_{\gamma g}(x)$ is a **multivalued set** and there may exist $\hat{x} \in \mathbb{R}^n$ s.t. $\text{prox}_{\gamma g}(\hat{x}) = \emptyset$.

The proximal operator: definition

Crucial tool for the development of non-smooth optimisation algorithms. Relations with activation functions in the context of deep networks (Combettes, Pesquet, '20).

Definition

Let $g \in \mathcal{P}$. Then, the *proximal operator* of g with parameter $\gamma > 0$ is defined as the **multi-valued map** $\text{prox}_{\gamma g} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ defined for all $x \in \mathbb{R}^n$:

$$\text{prox}_{\gamma g}(x) := \arg\min_{y \in \mathbb{R}^n} \underbrace{g(y) + \frac{1}{2\gamma}\|y - x\|^2}_{=:h(y;x)}$$

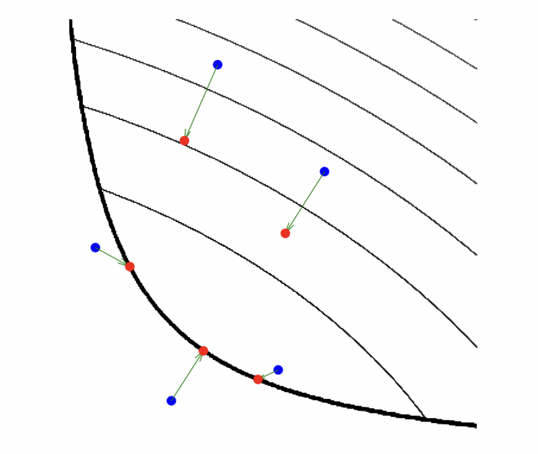
With no further conditions on g, $\text{prox}_{\gamma g}(x)$ is a **multivalued set** and there may exist $\hat{x} \in \mathbb{R}^n$ s.t. $\text{prox}_{\gamma g}(\hat{x}) = \emptyset$.

Proposition (uniqueness of the proximal point)

If $g \in \Gamma_0(\mathbb{R}^n)$, then $\text{prox}_{\gamma g}(x)$ exists and it is unique for all $x \in \mathbb{R}^n$.

"*Proof*": For all $x \in \mathbb{R}^n$, the function $h(\cdot; x)$ is $\frac{1}{\gamma}$-strongly (hence strictly) convex, hence it admits a unique minimiser.

Graphical interpretation



Thin black lines: level lines of g. **Thick** black lines: boundary of domain. Blue points: evaluation points are moved to the red points in the minimisation with an amount depending on γ. Note: points are moved to the minimum of the function.

For $\gamma > 0$ and $x \in \mathbb{R}^n$, let $z := \text{prox}_{\gamma g}(x)$. We have:

$$
\begin{aligned}
z := \text{prox}_{\gamma g}(x) \quad &\Leftrightarrow \quad z = \arg\min_{y \in \mathbb{R}^n} g(y) + \frac{1}{2\gamma}\|y - x\|^2 \\
\text{(optimality)} \quad &\Leftrightarrow \quad 0 \in \partial g(z) + \frac{1}{\gamma}(z - x) \\
\text{(rearranging)} \quad &\Leftrightarrow \quad x \in z + \gamma \partial g(z) \\
\text{(using operators)} \quad &\Leftrightarrow \quad x \in (Id + \gamma \partial g)(z) \\
\text{(uniqueness)} \quad &\Leftrightarrow \quad z = (Id + \gamma \partial g)^{-1}(x)
\end{aligned}
$$

[1] Minty, (1962), Bauschke-Combettes, (2010). Chambolle-Pock, (2016)

Relation with subdifferentials

For $\gamma > 0$ and $x \in \mathbb{R}^n$, let $z := \text{prox}_{\gamma g}(x)$. We have:

$$\begin{aligned}
z := \text{prox}_{\gamma g}(x) \quad &\Leftrightarrow \quad z = \arg\min_{y \in \mathbb{R}^n} g(y) + \frac{1}{2\gamma}\|y - x\|^2 \\
\text{(optimality)} \quad &\Leftrightarrow \quad 0 \in \partial g(z) + \frac{1}{\gamma}(z - x) \\
\text{(rearranging)} \quad &\Leftrightarrow \quad x \in z + \gamma \partial g(z) \\
\text{(using operators)} \quad &\Leftrightarrow \quad x \in (Id + \gamma \partial g)(z) \\
\text{(uniqueness)} \quad &\Leftrightarrow \quad z = (Id + \gamma \partial g)^{-1}(x)
\end{aligned}$$

For those of you who are familiar with convex analysis...

Remark[1]

$z = \text{prox}_{\gamma g}(x)$ is given by the *resolvent* of the maximal monotone operator $\gamma \partial g$ evaluated at x.

[1] Minty, (1962), Bauschke-Combettes, (2010). Chambolle-Pock, (2016)

Firm non-expansiveness of the proximal operator

Proposition (firm non-expansiveness)

Let $g \in \Gamma_0(\mathbb{R}^n)$. Then:

$$(\forall x \in \mathbb{R}^n) \quad \|\mathrm{prox}_g(x) - \mathrm{prox}_g(y)\|^2 \leq \langle x - y, \mathrm{prox}_g(x) - \mathrm{prox}_g(y)\rangle$$

Proof: There holds:

$$x - \mathrm{prox}_g(x) \in \partial f(\mathrm{prox}_g(x)), \quad y - \mathrm{prox}_g(y) \in \partial f(\mathrm{prox}_g(y)).$$

By definition of subdifferential:

$$f(\mathrm{prox}_g(y)) \geq f(\mathrm{prox}_g(x)) + \langle x - \mathrm{prox}_g(x), \mathrm{prox}_g(y) - \mathrm{prox}_g(x)\rangle,$$

and similarly inverting x and y. Summing:

$$\cancel{f(\mathrm{prox}_g(y))} + \cancel{f(\mathrm{prox}_g(x))}$$
$$\geq \cancel{f(\mathrm{prox}_g(y))} + \cancel{f(\mathrm{prox}_g(x))} + \langle y - f(\mathrm{prox}_g(y)) - x + f(\mathrm{prox}_g(x)), f(\mathrm{prox}_g(x)) - f(\mathrm{prox}_g(y))\rangle.$$

This implies non-expansiveness since:

$$\|\mathrm{prox}_g(x)-\mathrm{prox}_g(y)\|^{\cancel{2}} \leq \langle x-y, \mathrm{prox}_g(x)-\mathrm{prox}_g(y)\rangle \leq \|x-y\|\cancel{\|\mathrm{prox}_g(x) - \mathrm{prox}_g(y)\|}$$

Computation of proximal operators: indicator function

Example: Let $C \subset \mathbb{R}^n$ be a closed and convex set. Recall indicator function of C as:

$$\iota_C(x) := \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{if } x \notin C \end{cases}$$

The function $\iota_C(x)$ is proper, convex and l.s.c.

Computation of proximal operators: indicator function

Example: Let $C \subset \mathbb{R}^n$ be a closed and convex set. Recall indicator function of C as:

$$\iota_C(x) := \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{if } x \notin C \end{cases}$$

The function $\iota_C(x)$ is proper, convex and l.s.c.

$$\text{prox}_{\gamma \iota_C}(x) = \arg\min_{y \in \mathbb{R}^n} \; \iota_C(y) + \frac{1}{2\gamma}\|y - x\|^2 = \arg\min_{y \in C} \frac{1}{2\gamma}\|y - x\|^2 = P_C(x),$$

i.e. the **projection** of x onto C (the closest point $y \in C$ to x).

The notion of prox for functions g more general than ι_C is the reason why the prox operator is often referred to as *generalised projection*.

Computation of proximal operators: ℓ_1 norm

Example: Let $g(x) = |x|$ and $\gamma > 0$:

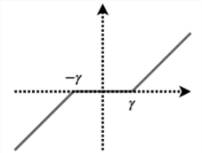
$$w = \text{prox}_{\gamma g}(x) = \underset{y \in \mathbb{R}}{\arg\min} \; |y| + \frac{1}{2\gamma}(y - x)^2$$

By optimality:

$$\gamma p + w - x = 0, \quad p \in \partial |w| \quad \Leftrightarrow \quad w = x - \gamma p, \quad p \in \partial |w|$$

Recalling the expression of $\partial | \cdot |$, one finds the definition of the *soft-thresholding* function

$$w = \text{prox}_{\gamma g}(x) = \begin{cases} x - \gamma & \text{if} \quad x > \gamma \\ x + \gamma & \text{if} \quad x < -\gamma \\ 0 & \text{if} \quad -\gamma \le x \le \gamma \end{cases} = \mathcal{T}_\gamma(x) := \text{sign}(x) \max\{|x| - \gamma, 0\}$$



A non-convex example: the ℓ_0 pseudo-norm

Example: Take

$$g(x) = \lambda |x|_0 := \begin{cases} \lambda & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

We want to compute:

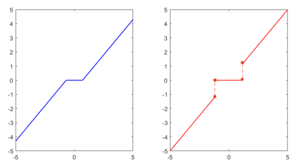
$$\mathrm{prox}_{\lambda|\cdot|_0}(z) = \arg\min_{y \in \mathbb{R}} \; h(y) := \frac{1}{2\lambda}(y - z)^2 + |y|_0$$

- if $y = 0$, then $h(0) = \frac{1}{2\lambda} z^2$
- if $y \neq 0$, then the minimum is reached at $y^* = z$, and $h(y^*) = 1$

By comparison we get:

$$h(0) = \frac{1}{2\lambda} z^2 \leq h(y^*) = 1 \Leftrightarrow z^2 \leq 2\lambda \Leftrightarrow -\sqrt{2\lambda} < z < \sqrt{2\lambda}$$

Therefore:



$$\mathcal{H}_{\sqrt{2\lambda}}(z) := \mathrm{prox}_{\lambda|\cdot|_0}(z) = \begin{cases} 0 & \text{if } |z| < \sqrt{2\lambda} \\ z & \text{if } |z| > \sqrt{2\lambda} \\ \{0, z\} & \text{if } |z| = \sqrt{2\lambda} \end{cases}$$

Soft VS. hard thresholding.

Computation of proximal points: properties

Proposition (proximal operator of separable functions)

Let $g \in \Gamma_0(\mathbb{R}^n)$ be **separable**, i.e. $g(x) = \sum_{i=1}^n g_i(x_i)$ for functions $g_i \in \Gamma_0(\mathbb{R})$. Then for $\gamma > 0$

$$\mathrm{prox}_{\gamma g}(x) = \left(\mathrm{prox}_{\gamma g_1}(x_1), \ldots, \mathrm{prox}_{\gamma g_n}(x_n)\right),$$

- $g(x) = \lambda \|x\|_1$, then $\mathrm{prox}_{\lambda\|\cdot\|_1}(x) = (\mathcal{T}_\lambda(x_i))_{i=1}^n = \mathcal{T}_\lambda(x)$.
- $g(x) = \lambda \|x\|_0$, then:

$$\mathrm{prox}_{\lambda\|\cdot\|_0} = \mathcal{H}_{\sqrt{2\lambda}}(x_1) \times \ldots \times \mathcal{H}_{\sqrt{2\lambda}}(x_n).$$

Proposition (proximal operators of rescaled and perturbed functions)

Let $g \in \Gamma_0(\mathbb{R}^n)$ and $\lambda \neq 0$. Define $h_1(x) := \lambda g(x/\lambda)$. Then, for $\gamma \in \mathbb{R}_{++}$:

$$\mathrm{prox}_{\gamma h_1}(x) = \lambda \mathrm{prox}_{\frac{\gamma}{\lambda} g}(x/\lambda).$$

Let $h_2(x) := \alpha g(x) + \frac{\beta}{2}\|x\|^2$, for $\alpha, \beta \in \mathbb{R}_{++}$. Then, for $\gamma \in \mathbb{R}_{++}$:

$$\mathrm{prox}_{\gamma h_2}(x) = \mathrm{prox}_{\frac{\alpha\gamma}{1+\beta\gamma} g}\left(\frac{x}{1+\beta\gamma}\right).$$

Let $h_3(x) := g(Wx)$ where $W \in \mathbb{R}^{m\times n}$ is orthogonal, $W^T W = Id$. Then, for $\gamma \in \mathbb{R}_{++}$:

$$\mathrm{prox}_{\gamma h_3}(x) = W^T \mathrm{prox}_{\gamma g}(Wx).$$

Computation of proximal points in general cases

Important remark

Having formulas for closed-form expressions of proximal points is very handy. Otherwise, a minimisation problem needs to be solved!

However, general regularisers **do not have this property**!

For more examples of easily-proximable function, see, e.g.:

- Beck, *First-order methods in optimization* 2006 (Chapter 6): many examples of proximal operators
- Parikh, Boyd, *Proximal algorithms*, 2013
- `http://proximity-operator.net/index.html`

In the lab class, we will make use of easily proximable (aka *simple*) functions. For non-proximable functions (e.g. TV) alternative strategies/algorithms should be found:

- Fenchel duality
- Smoothing
- Other algorithms (e.g., ADMM: Alessandro Lanza's computational imaging lab)

Non-smooth optimisation

Projected gradient descent

Towards forward-bacwkard splitting: projected gradient descent

For differentiable $f \in \Gamma_0(\mathbb{R}^n)$ and convex, closed $C \in \mathbb{R}^n$:

$$\arg\min_{x\in C} f(x) = \arg\min_{x\in\mathbb{R}^n} f(x) + \iota_C(x)$$

Algorithm: Projected Gradient Descent (PGD) algorithm

Input: $\tau \in \left(0, \frac{1}{L}\right], x^0 \in \mathbb{R}^n$.
for $k \geq 0$ **do**

$$x_{k+\frac{1}{2}} = x_k - \tau \nabla f(x_k)$$

$$x_{k+1} = P_C(x_{k+\frac{1}{2}}) = \arg\min_{y\in C} \frac{1}{2}\|y - x_{k+\frac{1}{2}}\|^2$$

$$= \arg\min_{y\in\mathbb{R}^n} \iota_C(y) + \frac{1}{2}\|y - x_{k+\frac{1}{2}}\|^2 = \operatorname{prox}_{\iota_C}(x_{k+\frac{1}{2}})$$

end for

- First: gradient step, next projection step
- Starting point for generalisation to more general convex, non-differentiable functions g...

Towards forward-backward splitting: explicit/implict GD

Let $f, g \in \Gamma_0(\mathbb{R}^n)$ and let f be smooth. Want to solve:

$$\arg\min_{x \in \mathbb{R}^n} f(x) + g(x)$$

Consider for $x_0 \in \mathbb{R}^n$, suitable $\tau > 0$ and $k \geq 0$, the following iterative scheme:

$$x_{k+1} \in x_k - \tau \nabla f(x_k) - \tau \partial g(x_{k+1}) \quad \Leftrightarrow \quad (Id + \tau \partial g(\cdot))(x_{k+1}) \in x_k - \tau \nabla f(x_k)$$

$$x_{k+1} \in (Id + \tau \partial g(\cdot))^{-1}(x_k - \tau \nabla f(x_k)) \quad \Leftrightarrow \quad x_{k+1} = \mathrm{prox}_{\tau g}(x_k - \tau \nabla f(x_k))$$

- Explicit GD on the smooth part f
- Implicit GD on the non-smooth part g

The proximal gradient algorithm

Framework: recap

$$\arg\min_{x\in\mathbb{R}^n} \left\{F(x) := f(x) + g(x)\right\},$$

- $f \in \Gamma_0(\mathbb{R}^n)$ is differentiable with L-Lipschitz continuous gradient

$$\exists L > 0, \quad (\forall x, y \in \mathbb{R}^n) \quad \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

- $g \in \Gamma_0(\mathbb{R}^n)$ is typically non-smooth but (assume) **easily-proximable!**

Examples: $g(x) = \iota_C(x)$, $g(x) = \|x\|_1$, $g(x) = \|x\|_1 + \iota_{\geq 0}(x)$, $g(x) = \|x\|_1 + \frac{\lambda}{2}\|x\|_2^2$, $g(x) = \|Wx\|_1$ with W orthogonal...

Algorithm: Forward-backward splitting (FB/FBS) algorithm[2]

Input: $x_0 \in \mathbb{R}^n$, $\tau \in \left(0, \frac{1}{L}\right]$.
for $k \geq 0$ **do**

$$x_{k+1} = \mathrm{prox}_{\tau g}\left(x_k - \tau\nabla f(x_k)\right)$$

end for

---

[2] Combettes, Wajs, 2005, Combettes, Pesquet, 2007

Remarks

- Step-size τ: still depending on the inverse of L, as for GD. If L is unknown/difficult to compute, **backtracking** strategies can be used, $\tau = \tau_k$ with suitable update rules.
- If g is easily proximable: no inner minimisation. Otherwise: need to solve a nested minimisation problem up to some accuracy (**inexact** algorithms).
- Computational cost/complexity: evaluation of ∇f may be costly (matrix/vector products), number of iterations before convergence depends on τ.
  - * Too small τ: unnecessary too many iterations
  - * Too big τ: risk of moving to a point z for which $F(z) > F(x_k)\dots$

Particular cases

- If $g \equiv 0$: smooth-optimisation problem. FBS reduces to GD.
- If $g(x) = \iota_C(x)$ for closed and convex $C \to$ PGD.
- If $g(x) = \lambda \|Wx\|_1$ for $\lambda > 0$ and orthogonal $W \in \mathbb{R}^{N \times n}$ (Wavelet basis...)

$$\min_{x \in \mathbb{R}^n} \; f(x) + \lambda \|Wx\|_1,$$

then the algorithm takes the structure of the Iterative Soft-Thresholding Algorithm (ISTA)

Iterative Soft Thresholding Algorithm (ISTA)[3]

The FB iteration takes the form:

$$x_{k+1} = W^T \mathcal{T}_{\tau\lambda}(Wx_k - \tau W \nabla f(x_k)),$$

where $\mathcal{T}_{\tau\lambda}(\cdot)$ is the *soft-thresholding* operator:

$$\mathcal{T}_{\tau\lambda}(z) = (\mathcal{T}_{\tau\lambda}(z_j))_{j=1,\dots,n} = \left(\left[|z_j| - \lambda\tau \right]_+ \operatorname{sign}(z_j) \right)_{j=1,\dots,n}$$

[3] Daubechies, Defrise, De Mol, 2004

The proximal gradient algorithm

Convergence properties

Convergence of FB iterations

Theorem (convergence of FB)[4]

Let $(x_k)_k$ the sequence of iterates generated by FB. Then, if $\tau \in (0, 1/L]$, there holds:

$$F(x_k) - F(x^*) \leq \frac{\|x^0 - x^*\|^2}{2\tau k}.$$

If, additionally, f or g are strongly convex with parameters $\mu_f, \mu_g > 0$ with $\mu := \mu_f + \mu_g$, then:

$$F(x_k) - F(x^*) + \frac{1 + \tau\mu_g}{2\tau}\|x_k - x^*\|^2 \leq \omega^k \frac{(1 + \tau\mu_g)\|x^0 - x^*\|^2}{2\tau},$$

with $\omega = \frac{1-\tau\mu_f}{1+\tau\mu_g} < 1$.

Same $O(1/k)/O(\omega^k)$ rates as for GD! Alternative way of seeing this: for $\epsilon > 0$, the iterates to get an ϵ-solution, i.e. x_k s.t.:

$$F(x_k) - F(x^*) \leq \epsilon$$

is $k \geq \lceil C/\epsilon \rceil$ and $k \geq \lceil C \log(1/\epsilon) \rceil$.

[4] Chambolle-Pock, 2016

Towards the proof: a generalised descent lemma

For all $k \geq$ and $\tau \in (0, 1/L]$ let:

$$x_{k+1} = T_\tau(x_k) := \text{prox}_{\tau g}(x_k - \tau \nabla f(x_k))$$

Generalised descent lemma

Let $\mu := \mu_f + \mu_g \geq 0$. Then, for all $x \in \mathbb{R}^n$, there holds:

$$F(x_{k+1}) + (1 + \tau\mu_g)\frac{\|x - x_{k+1}\|^2}{2\tau} \leq F(x) + (1 - \tau\mu_f)\frac{\|x - x_k\|^2}{2\tau}$$

Proof: By definition x_{k+1} solves:

$$x_{k+1} = \arg\min_x \; g(x) + f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{\|x - x_k\|^2}{2\tau}$$

By strong convexity there holds:

$$\overbrace{f(x) + g(x)}^{F(x)} + (1 - \tau\mu_f)\frac{\|x - x_k\|^2}{2\tau} \overbrace{\geq}^{\text{s.c. of } f} f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{\|x - x_k\|^2}{2\tau} + g(x)$$

$$\overbrace{\geq}^{\text{minimality and } \mu_g + \frac{1}{\tau} \text{ s.c.}} f(x_k) + g(x_{k+1}) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{\|x_{k+1} - x_k\|^2}{2\tau} + (1 + \tau\mu_g)\frac{\|x - x_{k+1}\|^2}{2\tau}$$

$\geq \dots$

Since f is L-Lipschitz there holds: $f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle \geq f(x_{k+1}) - \frac{L}{2}\|x_{k+1} - x_k\|^2$, hence:

$$\dots \geq F(x_{k+1}) + (1 + \tau\mu_g)\frac{\|x - x_{k+1}\|^2}{2\tau} + \underbrace{\left(\frac{1}{2\tau} - \frac{L}{2}\right)}_{\geq 0}\|x_{k+1} - x_k\|^2.$$

Convergence of FB: proof

Proof: Apply the generalised descent lemma for $x = x_k$, get:

$$F(x_{k+1}) \leq F(x_{k+1}) + (1+\tau\mu_g)\frac{\|x_k - x_{k+1}\|^2}{2\tau} \leq F(x_k),$$

so F is decreasing. Define $\omega := \frac{1-\tau\mu_f}{1+\tau\mu_g} \leq 1$, apply again the generalised descent lemma, which for $k = 0, \dots, K-1$ can be multiplied by ω^{-k-1} and summed:

$$\sum_{k=1}^{K} \omega^{-K}(F(x_k) - F(x)) + \sum_{k=1}^{K} \omega^{-k}\frac{1+\tau\mu_g}{2\tau}\|x - x_k\|^2 \leq \sum_{k=0}^{K-1} \omega^{-k-1}\frac{1-\tau\mu_f}{2\tau}\|x - x_k\|^2.$$

After cancellations, and using that $F(x_k) \geq F(x_K)$, for all $k = 0, \dots, K$, we get:

$$\omega^{-K}\left(\sum_{k=0}^{K-1}\omega^k\right)(F(x_K) - F(x)) + \omega^{-K}\frac{1+\tau\mu_g}{2\tau}\|x - x_K\|^2 \leq \frac{1+\tau\mu_g}{2\tau}\|x - x_0\|^2.$$

- $\mu = 0$, $\omega = 1$: we deduce the result observing that $\sum_{k=0}^{K-1}\omega^k = \sum_{k=0}^{K-1} 1 = K$.
- $\mu > 0$, $\omega < 1$: we deduce the linear rate by multiplying by ω^K and observing that $\sum_{k=0}^{K-1}\omega^k = \frac{1-\omega^K}{1-\omega} \geq 1$.

Analysis of the forward-backward algorithm: convergence of the sequence

We focus on the simple convex case (i.e. $\mu = 0$). For $\mu > 0$ this holds *a fortiori*.

Proposition (Fejér monotonicity)

Let (x_k) be the sequence generated by the FB algorithm with a constant stepsize $\tau \in (0, 1/L]$. Then, for any $x^* \in \arg\min\ F$, there holds:

$$\|x_{k+1} - x^*\| \leq \|x_k - x^*\|.$$

Lemma (convergence under Fejér monotonicity)

Let $(x_k) \subset \mathbb{R}^n$ be a sequence and let: $D := \{\tilde{x} : \tilde{x} \text{ is a limiting pont of } (x_k)\}$. Let S s.t. $D \subseteq S$. If (x_k) is Fejér monotone for all elements $x^* \in S$, then it converges to a point in D.

Analysis of the forward-backward algorithm: convergence of the sequence

We focus on the simple convex case (i.e. $\mu = 0$). For $\mu > 0$ this holds *a fortiori*.

Proposition (Fejér monotonicity)

Let (x_k) be the sequence generated by the FB algorithm with a constant stepsize $\tau \in (0, 1/L]$. Then, for any $x^* \in \arg\min\ F$, there holds:

$$\|x_{k+1} - x^*\| \leq \|x_k - x^*\|.$$

Lemma (convergence under Fejér monotonicity)

Let $(x_k) \subset \mathbb{R}^n$ be a sequence and let: $D := \{\tilde{x} : \tilde{x} \text{ is a limiting pont of } (x_k)\}$. Let S s.t. $D \subseteq S$. If (x_k) is Fejér monotone for all elements $x^* \in S$, then it converges to a point in D.

Theorem (convergence of the iterates of FB)

Let (x_k) be the sequence generated by the FB algorithm with a constant step-size $\tau \in (0, 1/L]$. Then, $x_k \to x^*$, where $x^* \in \arg\min F$.

Proof: Let $\tilde{x}$ be a limit point of (x_k). Then, there exists a subsequence (x_{k_j}) such that $x_{k_j} \to \tilde{x}$. Then, since

$$F(x_{k_j}) - F(x^*) \to 0, \quad \text{for } j \to +\infty.$$

and F is l.s.c., we deduce:

$$F(\tilde{x}) \leq \liminf_{j \to +\infty} F(x_{k_j}) = F(x^*).$$

By minimality, $\tilde{x} \in \arg\min\ F$. By now defining $S := \mathit{argmin}\ F$ and applying the Lemma the thesis follows since all limiting points are elements of S.

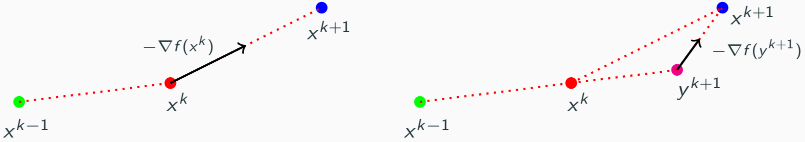
Acceleration strategies

Acceleration strategies

FISTA

Accelerated proximal gradient algorithm

Idea: add inertia to "shift" the sequence of iterates.



Algorithm: Fast Iterative Soft-Thresholding Algorithm (FISTA)[5]

Input: $x_0 = y_0 \in \mathbb{R}^n$, $\tau \in \left(0, \frac{1}{L}\right]$, $t_0 = 1$.
for $k \geq 0$ **do**

$$x_{k+1} = \operatorname{prox}_{\tau g}(y_k - \tau \nabla f(y_k))$$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$$

$$y_{k+1} = x_{k+1} + \frac{t_k - 1}{t_{k+1}}(x_{k+1} - x_k)$$

end for

[5] Nesterov, 2004 (APGD), Beck, Teboulle, 2009 (general g)

Properties of the parameter sequence

Proposition

Let $\{t_k\}$ be the sequence defined by $t_0 = 1$ and $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ for $k \geq 0$. Then:

$$t_k \geq \frac{k+2}{3} \quad \forall k \geq 0.$$

Proof: By induction. For $k = 0$:, obviously there holds: $t_0 = 1 \geq \frac{0+2}{2} = 1$. Suppose the claim holds for some $k > 0$. Using the recursion:

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2} \geq \frac{1+\sqrt{1+(k+2)^2}}{2} \geq \frac{1+\sqrt{(k+2)^2}}{2} = \frac{k+3}{2}.$$

Alternative choices: The sequence $\{t_k\}$ can alternatively be chosen so as to satisfy the following two properties holding for all $k \geq 0$:

- $t_k \geq \frac{k+2}{2}$
- $t_{k+1}^2 - t_{k+1} \leq t_k^2$.

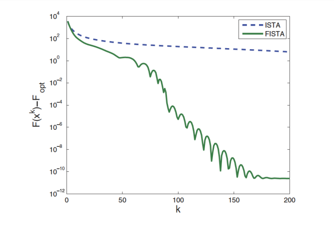
For instance, the choice $t_k = \frac{k+2}{2}$ satisfies both properties (Chambolle, Dossal, '15).

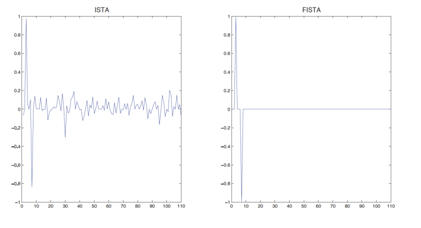
Theorem (Accelerated convergence of FISTA)

Let (x_k) the sequence of iterates generated by FISTA with $\tau \in (0, 1/L]$. Then, for any $x^* \in \arg\min\ F$, there holds:

$$F(x_k) - F(x^*) \leq \frac{2\|x_0 - x^*\|^2}{\tau(k+1)^2}$$

Proof: you will see this in the exercise class tomorrow with $\tau = 1/L$.





Accuracy viewpoint: w.r.t. to the vanilla FB algorithm, an ϵ-accurate solution, i.e.:

$$F(x_k) - F(x^*) \leq \epsilon$$

is obtained for $k \geq \lceil C/\sqrt{\epsilon} - 1 \rceil$.

Acceleration strategies

Strongly convex FISTA

A strongly convex variant of FISTA

Assume now that f is strongly convex with $\mu_f > 0$. Consider the algorithm:

Algorithm: Strongly convex FISTA - V-FISTA [6]

Input: $x_0 = y_0 \in \mathbb{R}^n$, $\tau = \frac{1}{L}$, and $\kappa := \frac{L}{\mu_f}$.
for $k \geq 0$ **do**

$$x_{k+1} = \text{prox}_{\frac{1}{L}g}(y_k - \frac{1}{L}\nabla f(y_k))$$
$$y_{k+1} = x_{k+1} + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)(x_{k+1} - x_k)$$

end for

Note: **constant** inertial parameter defined in terms of $\kappa \geq 1$.

...Both L and μ_f are required (difficult to estimate in practice)!

[6] Beck, '17, Chambolle, Pock '16, Calatroni, Chambolle, '19 (adaptive backtracking), Rebegoldi, Calatroni, '21 (variable scaling)

Convergence rates for strongly convex FISTA

Theorem (convergence of strongly convex FISTA[7])

Let (x_k) be the sequence of iterates generated by the strongly convex variant of the FISTA algorithm. Then, there holds:

$$F(x_k) - F(x^*) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \left(F(x_0) - F(x^*) + \frac{\mu_f}{2}\|x_0 - x^*\|^2\right),$$

Proof: you will see this in the exercise classes.

- In Chambolle, Pock, '16, Calatroni, Chambolle, '19, Rebegoldi, Calatroni '22: strongly convex variant of FISTA allowing strong convexity both in f and in g (better in g!)
- In Aujol, Dossal, Labarriere, Rondebierre, '21: FISTA algorithm under PL condition for f with an automatic estimate of the strong convexity parameter μ_f

[7] Beck, '17

The FISTA club

- **Convergence of iterates**: OK for FB (based on monotonicity arguments), proved for FISTA in Chambolle, Dossal, '15;
- **Monotone variants**: MFISTA (Beck, Teboulle, '09)
- **Non-Euclidean, inexact variants**:, Schmidt, Roux and Bach, '11, Villa, Salzo, Baldassarre, Verri, '13, Bonettini, Rebegoldi, Ruggiero, '19
- **Strongly convex, inexact and scaled**: SAGE-FISTA (Rebegoldi, Calatroni, '22)
- **Adaptive backtracking** for estimating τ 'on-the-fly': Scheinberg, Goldfarb, Bai, '14, Calatroni, Chambolle, '19, Florea, Vorobyov, '20
- **Restarting schemes**: heuristic (O'Donoghue, Candès, '15), rigorous (Alamo et al., '19, Aujol, Dossal, Labarriere, Rondepierre et al., '21)
- **ODE interpretation**: interpretation as discretised dynamical systems (with different inertial/friction/damping terms) Su, Boyd, Candès, '14, lot of works by Attouch, Cabot, Chbani, Peypouquet
- **Learned versions**: LISTA (Gregor, Le Cunn, 2010)
- Faster-FISTA, Adaptive FISTA...

Conclusions

We discussed the use of proximal-based algorithms for **convex** structured (smooth+non-smooth) optimisation problems in the form:

$$\arg\min_{x} \; f(x) + g(x)$$

- We revised basic tools of convex analysis for generalising derivatives to non-smooth functions
- We defined, characterised and looked at some fundamental properties of the proximal operator
- We defined the forward-backward (aka proximal gradient method) generalising the GD algorithm to the structured case and show a general convergence result for strongly convex functions
- We discussed acceleration strategies à la Nesterov: FISTA and its strongly covex variants

Extensions

Extensions

Inexact algorithms

Inexact proximal evaluations

$$p = \operatorname{prox}_g(a) \Leftrightarrow p = \operatorname{argmin}_x \left\{ \phi(x) := g(x) + \frac{1}{2}\|x - a\|^2 \right\} \Leftrightarrow p - a \in \partial g(p)$$

Inexact proximal evaluations

$$p = \text{prox}_g(a) \Leftrightarrow p = \text{argmin}_x \left\{ \phi(x) := g(x) + \frac{1}{2}\|x-a\|^2 \right\} \Leftrightarrow p - a \in \partial g(p)$$

There are various ways to relax this to incorporate **errors**[8]

- **Type 1 errors** : $\hat{p} \approx_1^{\varepsilon} p$ if

$$\hat{p} \in \varepsilon - \text{argmin}_x \ \phi(x) := \left\{ x' \in \mathbb{R}^n : \phi(x') \le \inf \phi(x) + \varepsilon \right\}$$

[8] Salzo, Villa, '12, Villa, Salzo, Baldassarre, Verri, '13

Inexact proximal evaluations

$$p = \mathrm{prox}_g(a) \Leftrightarrow p = \mathrm{argmin}_x \left\{ \phi(x) := g(x) + \frac{1}{2}\|x - a\|^2 \right\} \Leftrightarrow p - a \in \partial g(p)$$

There are various ways to relax this to incorporate **errors**[8]

- **Type 1 errors**: $\hat{p} \approx_1^{\varepsilon} p$ if

$$\hat{p} \in \varepsilon - \mathrm{argmin}_x \ \phi(x) := \{x' \in \mathbb{R}^n : \phi(x') \leq \inf \phi(x) + \varepsilon\}$$

- **Type 2 errors**: $\hat{p} \approx_2^{\varepsilon} p$ if

$$\hat{p} - a \in \partial_{\varepsilon^2} g(\hat{p}) = \left\{ u \in \mathbb{R}^n : g(x') \geq g(\hat{p}) + u^T(x' - \hat{p}) - \varepsilon^2 \ \forall x' \right\}$$

[8] Salzo, Villa, '12, Villa, Salzo, Baldassarre, Verri, '13

Inexact proximal evaluations

$$p = \mathrm{prox}_g(a) \Leftrightarrow p = \mathrm{argmin}_x \left\{ \phi(x) := g(x) + \frac{1}{2}\|x - a\|^2 \right\} \Leftrightarrow p - a \in \partial g(p)$$

There are various ways to relax this to incorporate **errors**[8]

- **Type 1 errors**: $\hat{p} \approx_1^\varepsilon p$ if

$$\hat{p} \in \varepsilon - \mathrm{argmin}_x\ \phi(x) := \left\{ x' \in \mathbb{R}^n : \phi(x') \leq \inf \phi(x) + \varepsilon \right\}$$

- **Type 2 errors** : $\hat{p} \approx_2^\varepsilon p$ if

$$\hat{p} - a \in \partial_{\varepsilon^2} g(\hat{p}) = \left\{ u \in \mathbb{R}^n : g(x') \geq g(\hat{p}) + u^T(x' - \hat{p}) - \varepsilon^2 \ \forall x' \right\}$$

- **Type 3 errors** : $\hat{p} \approx_3^\varepsilon p$ if $\hat{p} = \mathrm{prox}_g(a + e), \|e\| \leq \varepsilon$.

[8] Salzo, Villa, '12, Villa, Salzo, Baldassarre, Verri, '13

Inexact proximal evaluations

$$p = \text{prox}_g(a) \Leftrightarrow p = \text{argmin}_x \left\{ \phi(x) := g(x) + \frac{1}{2}\|x - a\|^2 \right\} \Leftrightarrow p - a \in \partial g(p)$$

There are various ways to relax this to incorporate **errors**[8]

- **Type 1 errors**: $\hat{p} \approx_1^{\varepsilon} p$ if

$$\hat{p} \in \varepsilon - \text{argmin}_x \ \phi(x) := \left\{ x' \in \mathbb{R}^n : \phi(x') \leq \inf \phi(x) + \varepsilon \right\}$$

- **Type 2 errors** : $\hat{p} \approx_2^{\varepsilon} p$ if

$$\hat{p} - a \in \partial_{\varepsilon^2} g(\hat{p}) = \left\{ u \in \mathbb{R}^n : g(x') \geq g(\hat{p}) + u^T(x' - \hat{p}) - \varepsilon^2 \ \forall x' \right\}$$

- **Type 3 errors**: $\hat{p} \approx_3^{\varepsilon} p$ if $\hat{p} = \text{prox}_g(a + e), \|e\| \leq \varepsilon$.

Theorem (convergence of inexact FISTA)

For $\tau \leq 1/L$, if $\varepsilon_k = O(1/k^q)$ with $q > 3/2$, then the sequence (x_k) of the accelerated inexact FB algorithm satisfies:

$$F(x_k) - F(x^*) = O\left(\frac{1}{k^2}\right)$$

[8] Salzo, Villa, '12, Villa, Salzo, Baldassarre, Verri, '13

Extensions

Backtracking strategies for FISTA

FISTA with monotone backtracking[9]

For f convex and differentiable, define the Bregman "distance""

$$D_f(x,y) := f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq 0, \qquad \forall x, y \in \mathbb{R}^n$$

Popular for **mirror descent** algorithms and regularisation of inverse problems (Burger, '16).

Algorithm: FISTA with non-decreasing backtracking

Input: $x_0 = y_0 \in \mathbb{R}^n$, $\tau_0 > 0$, $t_0 = 1$, $\rho \in (0,1)$.
for $k \geq 0$ **do**
  for $i = 0, 1, \ldots$ **repeat**

$$\tau_{k+1} = \rho^i \tau_k$$
$$x_{k+1} = \text{prox}_{\tau_{k+1} g}(y_k - \tau_{k+1}\nabla f(y_k))$$
$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$$
$$y_{k+1} = x_{k+1} + \frac{t_k - 1}{t_{k+1}}(x_{k+1} - x_k)$$

  until $D_f(x^{k+1}, y^{k+1}) \leq \|x^{k+1} - y^{k+1}\|^2 / 2\tau_{k+1}$
end for

[9] Beck, Teboulle, '09, Chambolle, Pock, '16

Convergence guarantee for FISTA with non-adaptive backtracking

Theorem (FISTA with non-adaptive backtracking)

Let (x_k) the sequence of iterates generated by FISTA with non-adaptive backtracking. Then, for any $x^* \in \arg\min\ F$, there holds:

$$F(x_k) - F(x^*) \leq \frac{2\|x_0 - x^*\|^2}{\tau\rho(k+1)^2}$$

- Basically the same rate as before, just depending on $\rho \in (0, 1)$
- Idea: start in an optimistic way $\tau_0 \gg 1$. If at any step $k \geq 1$ the step-size is too big, it will be decreased up to guarantee decay

Non-monotone FISTA backtracking

Algorithm: FISTA with adaptive backtracking

Input: $x_0 = y_0 \in \mathbb{R}^n$, $\tau_0 > 0$, $t_0 = 1$, $\rho \in (0,1)$, $\delta \in (0,1)$.
for $k \geq 0$ **do**

$$\tau_{k+1}^0 = \frac{\tau_k}{\delta}; \qquad (*)$$

for $i = 0, 1, \ldots$ **repeat**

$$\tau_{k+1} = \rho^i \tau_{k+1}^0$$

$$x_{k+1} = \mathrm{prox}_{\tau_{k+1} g}(y_k - \tau_{k+1} \nabla f(y_k))$$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$$

$$y_{k+1} = x_{k+1} + \frac{t_k - 1}{t_{k+1}}(x_{k+1} - x_k)$$

until $D_f(x^{k+1}, y^{k+1}) \leq \|x^{k+1} - y^{k+1}\|^2 / 2\tau_{k+1}$.
end for

- **Only difference**: **tentative step** where you try to increase the previous step-size.
- Practically, you may even add a max number of backtracking iterations $i_{\max} \approx 10$

Convergence guarantee for FISTA with adaptive backtracking)

Theorem (FISTA with adaptive backtracking[10])

Let (x_k) the sequence of iterates generated by FISTA with non-adaptive backtracking. Then, for any $x^* \in \arg\min\ F$, there holds:

$$F(x_k) - F(x^*) \leq \frac{2\bar{L}_k}{k^2}\|x^0 - x^*\|^2 \leq \frac{2L}{\rho k^2}\|x^0 - x^*\|^2$$

where $\sqrt{\bar{L}_k} := \frac{1}{\frac{1}{k}\sum_{i=1}^k \frac{1}{\sqrt{L_i}}}$, $L_i := 1/\tau_i$.

From standard harmonic/arithmetic mean inequalities:

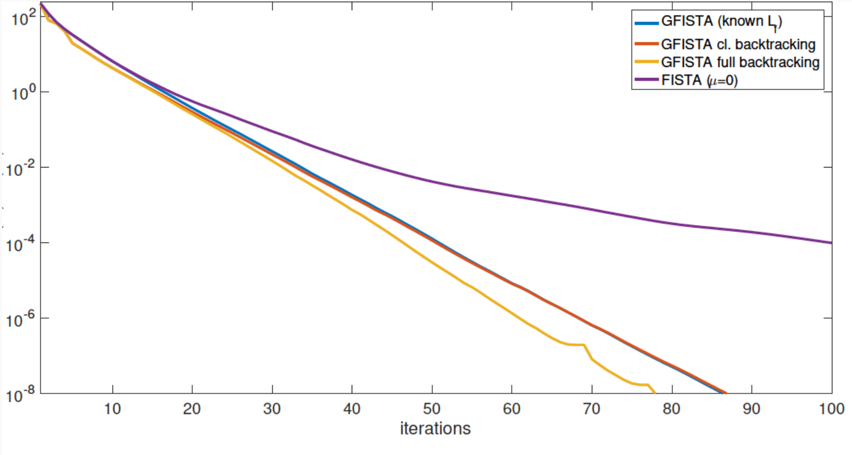
$$\sqrt{\bar{L}_k} \leq \frac{1}{k}\sum_{i=1}^k \sqrt{L_i} \leq \sqrt{\frac{1}{k}\sum_{i=1}^k L_i} \leq \sqrt{\frac{L}{\rho}}$$

- "Local" estimates: you don't need the dependence on L_f in final rates (which is in principle unknown), you have acceleration depending on harmonic mean
- Extensions in Rebegoldi, Calatroni' 22 to inexact proximal algorithms, with scaling.
- For step-size selection strategies in non-convex problems see Ochs, Chen, Brox, Pock, '14

[10] Scheinberg, Goldfarb, Bai, '14, Calatroni, Chambolle, '19

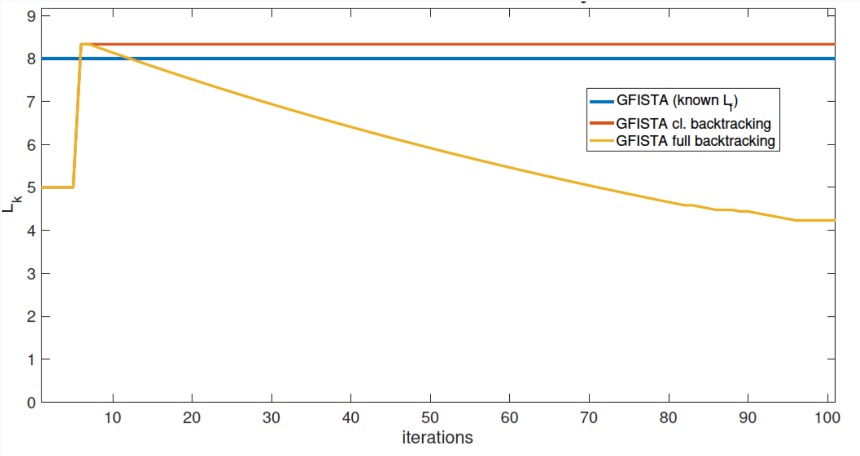
Backtracking performance

In Calatroni, Chambolle, '19 we considered a variation for strongly convex functions.



Backtracking performance

In Calatroni, Chambolle, '19 we considered a variation for strongly convex functions.



Non-convex algorithms

Gradient descent for non-convex functions

Let f be a C^2, L-smooth function which is coercive and bounded from below. Using Taylor expansion with integral form of remainder we have that:

$$
\begin{aligned}
f(x_{k+1}) &= f(x_k - \tau \nabla f(x_k)) \\
&= f(x_k) - \tau \langle \nabla f(x_k), \nabla f(x_k) \rangle + \int_0^\tau (\tau - t) \langle \nabla^2 f(x_k - t \nabla f(x_k)) \nabla f(x_k), \nabla f(x_k) \rangle dt \\
&\leq f(x_k) - \tau \left(1 - \frac{\tau L}{2}\right) \|\nabla f(x_k)\|^2
\end{aligned}
$$

as long as $\nabla^2 f \preceq L\mathrm{Id}$. Hence, if $\tau < 2/L$, the GD algorithm is decreasing and we can deduce that subsequences of (x_k) converge to some critical point.

A glimpse on the use of proximal gradient methods for non-convex problems

Theorem (Convergence of FB for non-convex f)

Let f be proper and L-smooth and $g \in \Gamma_0(\mathbb{R}^n)$. Let *argmin* $F \neq \emptyset$. Let (x_k) be the sequence generated by the FB algorithm with a constant stepsize $\bar{L} \in \left(\frac{L}{2}, +\infty\right)$. Then:

- the sequence $(F(x_k))$ is non-increasing and $F(x_{k+1}) < F(x_k)$ if and only if x_k is not a stationary point;
- The (generalised) gradient mapping $G_L : \text{int}(dom(f)) \to \mathbb{R}^n$ defined by:

$$G_{\bar{L}}(x) := \bar{L}\left(x - \text{prox}_{\frac{1}{\bar{L}}g}\left(x - \frac{1}{\bar{L}}\nabla f(x)\right)\right)$$

  is such that $G_{\bar{L}}(x_k) \to 0$ as $k \to +\infty$
- All limiting points of (x_k) are stationary points for the functional F.

- Earlier works by Fukushima, Mine, '81, Chouzenoux, Pesquet, Repetti, '14, Bredies, Lorenz, Reiterer, '15, Nesterov, '13.
- For results on accelerated algorithms see, e.g., Ochs, Chen, Brox, Pock, '14
- General convergence theory under the (non-restrictive) Kurdyka-Łojasiewicz property (Bolte, Daniilidis, Lewis, '06, Attouch, Bolte, Svaiter, '13, Attouch, Bolte, Redont, Subeyran, '14)

Questions?

calatroni@i3s.unice.fr