# First Order Descent Methods in Nonsmooth Convex Optimization

**Juan PEYPOUQUET**
Universidad Técnica Federico Santa María
Universidad de Chile

CIMPA School on Algorithmic Nonsmooth Optimization

Ibarra, September 19-21, 2017

# NONSMOOTH CONVEX FUNCTIONS AND SUBDIFFERENTIAL CALCULUS

# Extended real-valued functions

We consider $f : H \to \mathbb{R} \cup \{+\infty\}$[1], and write

$$\mathrm{dom}(f) = \{x \in H : f(x) < +\infty\}.$$

The indicator function of a set $C \subset H$ is

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

Then,

$$\min\{f(x) : x \in C\} = \min\{f(x) + \delta_C(x) : x \in H\}.$$

[1] We exclude here the case $f \equiv +\infty$.

# Extended real-valued functions

We consider $f : H \to \mathbb{R} \cup \{+\infty\}$[1], and write

$$\mathrm{dom}(f) = \{x \in H : f(x) < +\infty\}.$$

The indicator function of a set $C \subset H$ is

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

Then,

$$\min\{f(x) : x \in C\} = \min\{f(x) + \delta_C(x) : x \in H\}.$$

[1] We exclude here the case $f \equiv +\infty$.

# Extended real-valued functions

We consider $f : H \to \mathbb{R} \cup \{+\infty\}$[1], and write

$$\mathrm{dom}(f) = \{x \in H : f(x) < +\infty\}.$$

The indicator function of a set $C \subset H$ is

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

Then,

$$\min\{f(x) : x \in C\} = \min\{f(x) + \delta_C(x) : x \in H\}.$$

[1] We exclude here the case $f \equiv +\infty$.

# Convex functions

A function $f : H \to \mathbb{R} \cup \{+\infty\}$ is convex if

$$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y)$$

for every $x, y \in H$ and every $\lambda \in (0,1)$.

This is equivalent to the convexity of the epigraph of $f$:

$$\mathrm{epi}(f) = \{(x,z) \in H \times \mathbb{R} : f(x) \leq z\},$$

and implies the convexity of $\mathrm{dom}(f)$ and of every sublevel set

$$[f \leq \gamma] = \{x \in H : f(x) \leq \gamma\}, \qquad \gamma \in \mathbb{R}.$$

# Convex functions

A function $f : H \to \mathbb{R} \cup \{+\infty\}$ is convex if

$$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y)$$

for every $x, y \in H$ and every $\lambda \in (0,1)$.

This is equivalent to the convexity of the epigraph of $f$:

$$\mathrm{epi}(f) = \{(x,z) \in H \times \mathbb{R} : f(x) \leq z\},$$

and implies the convexity of $\mathrm{dom}(f)$ and of every sublevel set

$$[f \leq \gamma] = \{x \in H : f(x) \leq \gamma\}, \qquad \gamma \in \mathbb{R}.$$

# Lower-semicontinuity

A function $f : H \to \mathbb{R} \cup \{+\infty\}$ is lower-semicontinuous at $x \in \mathrm{dom}(f)$ if, for every $\varepsilon > 0$, there is $\delta > 0$ such that

$$\|x - y\| < \delta \Longrightarrow f(y) > f(x) - \varepsilon.$$

Equivalently, if $x_k \to x \Longrightarrow f(x) \leq \liminf_{k \to +\infty} f(x_k)$.

We say $f$ is lower-semicontinuous if it is so at every $x \in \mathrm{dom}(f)$.

**Example**

- *Every continuous function is lower-semicontinuous.*
- *$\delta_C$ is lower-semicontinuous if, and only if, $C$ is closed.*

# Lower-semicontinuity

A function $f : H \to \mathbb{R} \cup \{+\infty\}$ is lower-semicontinuous at $x \in \mathrm{dom}(f)$ if, for every $\varepsilon > 0$, there is $\delta > 0$ such that

$$\|x - y\| < \delta \Longrightarrow f(y) > f(x) - \varepsilon.$$

Equivalently, if $x_k \to x \Longrightarrow f(x) \leq \liminf_{k \to +\infty} f(x_k)$.

We say $f$ is lower-semicontinuous if it is so at every $x \in \mathrm{dom}(f)$.

**Example**

- *Every continuous function is lower-semicontinuous.*
- *$\delta_C$ is lower-semicontinuous if, and only if, $C$ is closed.*

# Lower-semicontinuity

A function $f : H \to \mathbb{R} \cup \{+\infty\}$ is lower-semicontinuous at $x \in \mathrm{dom}(f)$ if, for every $\varepsilon > 0$, there is $\delta > 0$ such that

$$\|x - y\| < \delta \Longrightarrow f(y) > f(x) - \varepsilon.$$

Equivalently, if $x_k \to x \Longrightarrow f(x) \leq \liminf_{k \to +\infty} f(x_k)$.

We say $f$ is lower-semicontinuous if it is so at every $x \in \mathrm{dom}(f)$.

## Example

- *Every continuous function is lower-semicontinuous.*
- $\delta_C$ *is lower-semicontinuous if, and only if, C is closed.*

# Two important properties

## Proposition

*Let $f : H \to \mathbb{R} \cup \{+\infty\}$. The following are equivalent:*

- *$f$ is lower-semicontinuous.*
- *$epi(f)$ is closed.*
- *$[f \leq \gamma]$ is closed for every $\gamma \in \mathbb{R}$.*

## Proposition

*Let $f : H \to \mathbb{R} \cup \{+\infty\}$ be convex and lower-semicontinuous. For every $x \in dom(f)$,*

$$x_k \rightharpoonup x \Longrightarrow f(x) \leq \liminf_{k \to +\infty} f(x_k).$$

# Two important properties

## Proposition

*Let $f : H \to \mathbb{R} \cup \{+\infty\}$. The following are equivalent:*

- *$f$ is lower-semicontinuous.*
- *epi($f$) is closed.*
- *$[f \leq \gamma]$ is closed for every $\gamma \in \mathbb{R}$.*

## Proposition

*Let $f : H \to \mathbb{R} \cup \{+\infty\}$ be convex and lower-semicontinuous. For every $x \in dom(f)$,*

$$x_k \rightharpoonup x \Longrightarrow f(x) \leq \liminf_{k \to +\infty} f(x_k).$$

# Existence of minimizers

A function $f : H \to \mathbb{R} \cup \{+\infty\}$ is coercive if

$$\lim_{\|x\| \to +\infty} f(x) = +\infty.$$

In other words, if every sublevel set $[f \leq \gamma]$ is bounded.

**Theorem (Weierstrass-Hilbert-Tonelli)**

*If $f : H \to \mathbb{R} \cup \{+\infty\}$ is convex, lower-semicontinuous and coercive, then $S \neq \emptyset$.*

# Existence of minimizers

A function $f : H \to \mathbb{R} \cup \{+\infty\}$ is coercive if

$$\lim_{\|x\| \to +\infty} f(x) = +\infty.$$

In other words, if every sublevel set $[f \leq \gamma]$ is bounded.

**Theorem (Weierstrass-Hilbert-Tonelli)**

*If $f : H \to \mathbb{R} \cup \{+\infty\}$ is convex, lower-semicontinuous and coercive, then $S \neq \emptyset$.*

# Existence of minimizers

A function $f : H \to \mathbb{R} \cup \{+\infty\}$ is coercive if

$$\lim_{\|x\| \to +\infty} f(x) = +\infty.$$

In other words, if every sublevel set $[f \leq \gamma]$ is bounded.

**Theorem (Weierstrass-Hilbert-Tonelli)**

*If $f : H \to \mathbb{R} \cup \{+\infty\}$ is convex, lower-semicontinuous and coercive, then $S \neq \emptyset$.*

# Subdifferential calculus

Recall that if $f : H \to \mathbb{R}$ is convex and differentiable, then

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \qquad \forall x, y \in H.$$

Keeping this in mind, we define the subdifferential of $f : H \to \mathbb{R} \cup \{+\infty\}$ at $x \in H$ by

$$\partial f(x) = \{x^* \in H : f(y) \geq f(x) + \langle x^*, y - x \rangle \quad \forall y \in H\}.$$

We write $\mathrm{dom}(\partial f) = \{x \in H : \partial f(x) \neq \emptyset\}$.

# Subdifferential calculus

Recall that if $f : H \to \mathbb{R}$ is convex and differentiable, then

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \qquad \forall x, y \in H.$$

Keeping this in mind, we define the subdifferential of $f : H \to \mathbb{R} \cup \{+\infty\}$ at $x \in H$ by

$$\partial f(x) = \{x^* \in H : f(y) \geq f(x) + \langle x^*, y - x \rangle \quad \forall y \in H\}.$$

We write $\mathrm{dom}(\partial f) = \{x \in H : \partial f(x) \neq \emptyset\}$.

# Subdifferential calculus

Recall that if $f : H \to \mathbb{R}$ is convex and differentiable, then

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \qquad \forall x, y \in H.$$

Keeping this in mind, we define the subdifferential of $f : H \to \mathbb{R} \cup \{+\infty\}$ at $x \in H$ by

$$\partial f(x) = \{x^* \in H : f(y) \geq f(x) + \langle x^*, y - x \rangle \quad \forall y \in H\}.$$

We write $\mathrm{dom}(\partial f) = \{x \in H : \partial f(x) \neq \emptyset\}$.

# Some properties

**Proposition**

*If $f$ is differentiable at $x$, then $\partial f(x) = \{\nabla f(x)\}$.*

**Proposition**

*If $x^* \in \partial f(x)$ and $y^* \in \partial f(y)$, then $\langle x^* - y^*, x - y \rangle \geq 0$.*

**Proposition (Fermat's Rule)**

$$x \in S \Longleftrightarrow 0 \in \partial f(x).$$

# Some properties

**Proposition**

*If $f$ is differentiable at $x$, then $\partial f(x) = \{\nabla f(x)\}$.*

**Proposition**

*If $x^* \in \partial f(x)$ and $y^* \in \partial f(y)$, then $\langle x^* - y^*, x - y \rangle \geq 0$.*

**Proposition (Fermat's Rule)**

$$x \in S \Longleftrightarrow 0 \in \partial f(x).$$

# Some properties

**Proposition**

*If $f$ is differentiable at $x$, then $\partial f(x) = \{\nabla f(x)\}$.*

**Proposition**

*If $x^* \in \partial f(x)$ and $y^* \in \partial f(y)$, then $\langle x^* - y^*, x - y \rangle \geq 0$.*

**Proposition (Fermat's Rule)**

$$x \in S \Longleftrightarrow 0 \in \partial f(x).$$

# A calculus rule

## Proposition (Moreau-Rockafellar)

*For $f, g : H \to \mathbb{R} \cup \{+\infty\}$, we have*

$$\partial f(x) + \partial g(x) \subseteq \partial(f+g)(x) \qquad \forall x \in H.$$

*If $g$ is continuous at some $\bar{x} \in dom(f)$, then*

$$\partial(f+g)(x) = \partial f(x) + \partial g(x) \qquad \forall x \in H.$$

Example: The version of Fermat's Rule we saw yesterday.

# MINIMIZATION OF NONSMOOTH FUNCTIONS

# The proximal point algorithm

Let $f : H \to \mathbb{R} \cup \{+\infty\}$ be convex and lower-semicontinuous. Given $z_k \in H$ y $\lambda > 0$, we define $z_{k+1}$ as the unique solution of

$$\operatorname{argmin} \left\{ f(z) + \frac{1}{2\lambda} \|z - z_k\|^2 \right\}.$$

**Remark**

- *It is well defined.*
- *There is no need for f to be differentiable (not even continuous).*

# The proximal point algorithm

Let $f : H \to \mathbb{R} \cup \{+\infty\}$ be convex and lower-semicontinuous. Given $z_k \in H$ y $\lambda > 0$, we define $z_{k+1}$ as the unique solution of

$$\operatorname{argmin} \left\{ f(z) + \frac{1}{2\lambda} \|z - z_k\|^2 \right\}.$$

**Remark**

- *It is well defined.*
- *There is no need for f to be differentiable (not even continuous).*

# Relationship with the previous systems

Steepest descent dynamics

$$\dot{x}(t) = -\nabla f(x(t))$$

Gradient method:

$$\frac{x_{k+1} - x_k}{\lambda} = -\nabla f(x_k) \qquad \Leftrightarrow \qquad x_{k+1} = x_k - \lambda \nabla f(x_k).$$

Proximal point algorithm:

$$\frac{z_{k+1} - z_k}{\lambda} \in -\partial f(z_{k+1}) \qquad \Leftrightarrow \qquad z_{k+1} + \lambda \partial f(z_{k+1}) \ni z_k.$$

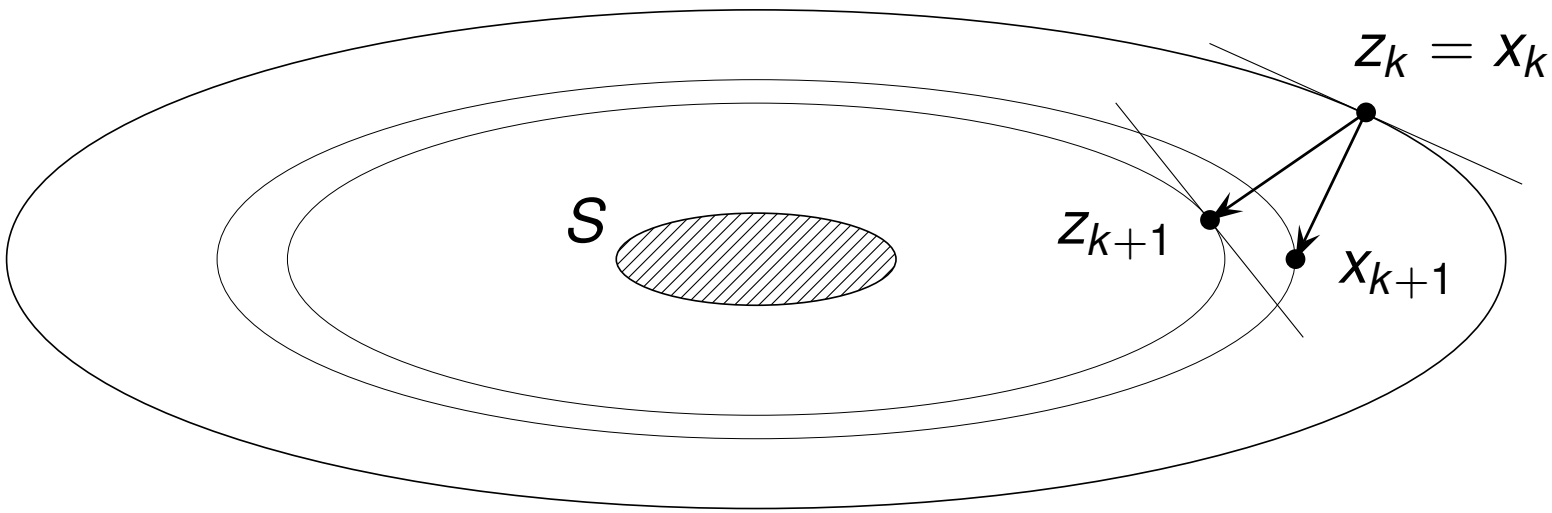
# Geometric comparison

Gradient

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

Proximal

$z_{k+1} + \lambda \nabla f(z_{k+1}) = z_k$



# Convergence

- The sequence $\big(f(z_k)\big)$ is nonincreasing.
- If $f$ is bounded from below, then $\sum_{k\geq 0} \|z_{k+1} - z_k\|^2 < +\infty$.
- $\lim_{k\to+\infty} f(z_k) = \inf(f)$. In particular, if $z_{n_k} \rightharpoonup \bar{z}$, then $\bar{z} \in S$.
- If $S \neq \emptyset$, then $f(z_k) - \min(f) \leq \dfrac{\operatorname{dist}(z_0, S)^2}{2\lambda k}$.
- For each $z^* \in S$, the sequence $\big(\|z_k - z^*\|\big)$ is nonincreasing.
- Let $S \neq \emptyset$. As $k \to +\infty$, $z_k$ converges weakly to a point in $S$.

# THE "SMOOTH + NONSMOOTH" STRUCTURE

# We have two basic methods

Gradient method (for a smooth $g$):

$$x_{k+1} = \mathrm{Grad}_{\lambda g}(x_k) = x_k - \lambda \nabla g(x_k)$$

Proximal point algorithm (for any $h$):

$$z_{k+1} = \mathrm{Prox}_{\lambda h}(z_k) = \mathrm{argmin}\left\{ h(z) + \frac{1}{2\lambda}\|z - z_k\|^2 \right\}$$

# The good and the bad

Gradient method

- \+ Easy to implement (explicit formula).
- – For convergence, $f$ is required to be smooth. The step size $\lambda$ must be properly chosen.

Proximal point algorithm

- \+ Convergence does not rely on the differentiability of $f$ or the step size.
- – The implementation may be difficult or costly (the formula is implcit).

# The good and the bad

Gradient method

- $+$ Easy to implement (explicit formula).
- $-$ For convergence, $f$ is required to be smooth. The step size $\lambda$ must be properly chosen.

Proximal point algorithm

- $+$ Convergence does not rely on the differentiability of $f$ or the step size.
- $-$ The implementation may be difficult or costly (the formula is implcit).

# Mixed structure

We want to minimize $f = g + h$, where $g$ is smooth but $h$ is not.

**Example**

*The constrained problem*

$$\min\{g(x) : x \in C\}$$

*is equivalent to*

$$\min\{g(x) + \delta_C(x) : x \in C\}.$$

# Mixed structure

We want to minimize $f = g + h$, where $g$ is smooth but $h$ is not.

## Example

*The constrained problem*

$$\min\{g(x) : x \in C\}$$

*is equivalent to*

$$\min\{g(x) + \delta_C(x) : x \in C\}.$$

# Mixed structure

We want to minimize $f = g + h$, where $g$ is smooth while $h$ is not.

## Example

*Let $\mu > 0$, $b \in \mathbb{R}^M$ and $A$ a matrix of size $M \times N$. Consider the functions $g, h : \mathbb{R}^N \to \mathbb{R}$ defined by*

$$g(x) = \mu \sum_{n=1}^{N} |x_n| \qquad \text{and} \qquad h(x) = \frac{1}{2}\|Ax - b\|^2$$

*The function $f = g + h$ appears in image processing, data analysis, among other contexts.*

# Strategy

Use the gradient method for the smooth part, and the proximal point algorithm for the nonsmooth one:

The proximal-gradient method:

$$x_{k+1} = T_\lambda(x_k) = \mathrm{Prox}_{\lambda h}\big(\mathrm{Grad}_{\lambda g}(x_k)\big)$$

Case $h = 0$: gradient

Case $h = \delta_C$: projected gradient $x_{k+1} = \mathrm{Proj}_C\big(\mathrm{Grad}_{\lambda g}(x_k)\big)$

Case $g = 0$: proximal

# Strategy

Use the gradient method for the smooth part, and the proximal point algorithm for the nonsmooth one:

The proximal-gradient method:

$$x_{k+1} = T_\lambda(x_k) = \mathrm{Prox}_{\lambda h}\big(\mathrm{Grad}_{\lambda g}(x_k)\big)$$

Case $h = 0$: gradient

Case $h = \delta_C$: projected gradient $x_{k+1} = \mathrm{Proj}_C\big(\mathrm{Grad}_{\lambda g}(x_k)\big)$

Case $g = 0$: proximal

# A fundamental property

## Proposition

*If $\nabla g$ is Lipschitz-continuous with constant $L$ and $\lambda L \leq 1$, then*

$$f(T_\lambda(x)) + \frac{\|y - T_\lambda(x)\|^2}{2\lambda} \leq f(y) + \frac{\|y - x\|^2}{2\lambda}.$$

# Convergence

- The sequence $\big(f(x_k)\big)$ is nonincreasing.
- If $f$ is bounded from below, then $\sum_{k\geq 0} \|x_{k+1} - x_k\|^2 < +\infty$.
- $\lim_{k\to+\infty} f(x_k) = \inf(f)$. In particular, if $x_{n_k} \rightharpoonup \bar{x}$, then $\bar{x} \in S$.
- If $S \neq \emptyset$, then $f(x_k) - \min(f) \leq \dfrac{\operatorname{dist}(x_0, S)^2}{2\lambda k}$.
- For each $x^* \in S$, the sequence $\big(\|x_k - x^*\|\big)$ is nonincreasing.
- Let $S \neq \emptyset$. As $k \to +\infty$, $x_k$ converges weakly to a point in $S$.

# EXERCISES

# Proximal-gradient method and $\ell^1 + \ell^2$ minimization

- For $g : \mathbb{R}^N \to \mathbb{R}$ defined by $g(x) = \frac{1}{2}\|Ax - b\|^2$, compute

$$\mathrm{Grad}_{\lambda g}(y) = y - \lambda \nabla g(y).$$

- For $h : \mathbb{R}^N \to \mathbb{R}$ defined by $h(x) = \mu\|x\|_1$, compute[2]

$$\mathrm{Prox}_{\lambda h}(y) = \mathrm{argmin}\left\{ h(x) + \frac{1}{2\lambda}\|x - y\|^2 \right\}.$$

- Write down the corresponding proximal-gradient iteration $x_{k+1} = T_\lambda(x_k)$ explicitly.

[2] Study the case $N = 1$ first.