Proximal-Based Methods

Tutorial

Amir Beck

Technion - Israel Institute of Technology
Haifa, Israel

Tutorial Overview

The tutorial is all about first order methods, specifically those based on proximal computations

- Background: extended real-valued functions, subgradients, conjugate functions, the proximal operator
- proximal gradient
- fast proximal gradient (FISTA)
- smoothing
- block proximal gradient
- dual proximal gradient

Complement of Tutorial Overview

Unfortunately, the following important topics are not included:

- primal and dual projected subgradient
- non-Euclidean algorithms (mirror descent, non-Euclidean proximal gradient)
- conditional gradient
- alternating minimization
- ADMM

Underlying Spaces

- We will assume that the underlying vector spaces, usually denoted by $\mathbb{V}$ or $\mathbb{E}$, are finite dimensional real inner product spaces with endowed inner product $\langle \cdot, \cdot \rangle$ and endowed norm $\| \cdot \|$.

Euclidean space: a finite dimensional real vector space equipped with an inner product $\langle \cdot, \cdot \rangle$ endowed with the norm $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$, which is also called the Euclidean norm.

- Except for one case, we will always assume that the underlying vector space is Euclidean

Extended Real-Valued Functions

- D. P. Bertsekas, A. Nedic and A. E. Ozdaglar, *Convex analysis and optimization* (2013).
- R. T. Rockafellar, *Convex analysis* (1970).

Extended Real-Valued Functions

- An extended real-valued function is a function defined over the entire underlying space that can take any real value, as well as the infinite values $-\infty$ and ∞.
- **Infinite values arithmetic:**

$$\begin{aligned} a+\infty=\infty+a &= \infty && (-\infty < a < \infty),\\ a-\infty=-\infty+a &= -\infty && (-\infty < a < \infty),\\ a\cdot\infty=\infty\cdot a &= \infty && (0 < a < \infty),\\ a\cdot(-\infty)=(-\infty)\cdot a &= -\infty && (0 < a < \infty),\\ a\cdot\infty=\infty\cdot a &= -\infty && (-\infty < a < 0),\\ a\cdot(-\infty)=(-\infty)\cdot a &= \infty && (-\infty < a < 0),\\ 0\cdot\infty=\infty\cdot 0=0\cdot(-\infty)=(-\infty)\cdot 0 &= 0. \end{aligned}$$

- For an extended real-valued function $f:\mathbb{E}\to[-\infty,\infty]$, the effective domain or just the domain is the set

$$\text{dom}(f)=\{\mathbf{x}\in\mathbb{E}: f(\mathbf{x})<\infty\}.$$

- For any subset $C\subseteq\mathbb{E}$, the indicator function of C is

$$\delta_C(\mathbf{x})=\begin{cases} 0 & \mathbf{x}\in C,\\ \infty & \mathbf{x}\notin C.\end{cases}$$

Closedness

- The epigraph of an extended real-valued function $f : \mathbb{E} \to [-\infty, \infty]$ is defined by
$$\text{epi}(f) = \{(\mathbf{x}, y) : f(\mathbf{x}) \leq y, \mathbf{x} \in \mathbb{E}, y \in \mathbb{R}\} \subseteq \mathbb{E} \times \mathbb{R}.$$
- A function $f : \mathbb{E} \to [-\infty, \infty]$ is called proper if it does not attain the value $-\infty$ and there exists at least one $\hat{\mathbf{x}} \in \mathbb{E}$ such that $f(\hat{\mathbf{x}}) < \infty$, meaning that $\text{dom}(f) \neq \emptyset$.
- A function $f : \mathbb{E} \to [-\infty, \infty]$ is called closed if its epigraph is closed.

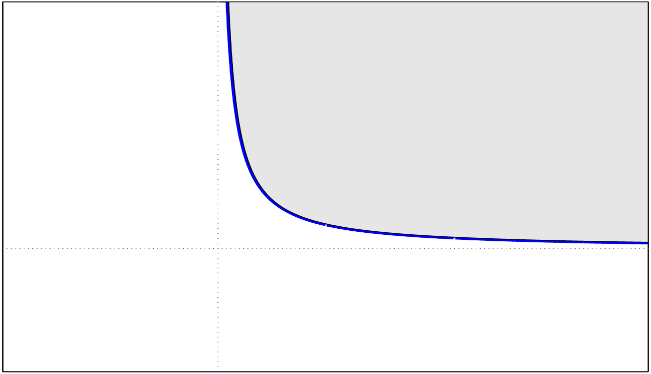
Theorem. The indicator function δ_C is closed if and only if C is closed.

Proof.
$$\text{epi}(f) = \{(\mathbf{x}, y) \in \mathbb{E} \times \mathbb{R} : \delta_C(\mathbf{x}) \leq y\} = C \times \mathbb{R}_+,$$
which is evidently closed if and only if C is closed. □

Example

$$f(x) = \begin{cases} \frac{1}{x}, & x > 0, \\ \infty, & \text{else.} \end{cases}$$

f is closed.



Lower Semicontinuity

Definition

- A function $f : \mathbb{E} \to [-\infty, \infty]$ is called lower semicontinuous at $\mathbf{x} \in \mathbb{E}$ if

$$f(\mathbf{x}) \leq \liminf_{n \to \infty} f(\mathbf{x}_n),$$

for any sequence $\{\mathbf{x}_n\}_{n \geq 1} \subseteq \mathbb{E}$ for which $\mathbf{x}_n \to \mathbf{x}$ as $n \to \infty$.

- A function $f : \mathbb{E} \to [-\infty, \infty]$ is called lower semicontinuous if it is lower semicontinuous at each point in $\mathbb{E}$.

Theorem. The following claims are equivalent:

(i) f is lower semicontinuous.

(ii) f is closed.

(iii) for any $\alpha \in \mathbb{R}$, the level set

$$\mathrm{Lev}(f, \alpha) = \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \leq \alpha\}$$

is closed.

Operations Preserving Closedness

Theorem.

(a) Let $\mathcal{A} : \mathbb{E} \to \mathbb{V}$ be a linear transformation and $\mathbf{b} \in \mathbb{V}$, and let $f : \mathbb{V} \to (-\infty, \infty]$ be closed. Then the function $g : \mathbb{E} \to [-\infty, \infty]$ given by

$$g(\mathbf{x}) = f(\mathcal{A}(\mathbf{x}) + \mathbf{b})$$

is closed.

(b) Let $f_1, f_2, \ldots, f_m : \mathbb{E} \to (-\infty, \infty]$ be extended real-valued closed functions, and let $\alpha_1, \alpha_2, \ldots, \alpha_m \in \mathbb{R}_+$. Then the function $f = \sum_{i=1}^{m} \alpha_i f_i$ is closed.

(c) Let $f_i : \mathbb{E} \to (-\infty, \infty], i \in I$ be extended real-valued closed functions, where I is a given index set. Then the function

$$f(\mathbf{x}) = \max_{i \in I} f_i(\mathbf{x}).$$

is closed.

Weierstrass theorem for closed functions

Theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper closed function, and assume that C is a compact set satisfying $C \cap \mathrm{dom}(f) \neq \emptyset$. Then

(a) f is bounded below over C.

(b) f attains a minimizer over C.

- A proper function $f : \mathbb{E} \to (-\infty, \infty]$ is called coercive if

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

Theorem. **(attainment under coerciveness)** Let $f : \mathbb{E} \to (-\infty, \infty]$ be a closed proper and coercive function and let $S \subseteq \mathbb{E}$ be a nonempty closed set satisfying $S \cap \mathrm{dom}(f) \neq \emptyset$. Then f attains a minimizer over S.

Convex Extended Real-Valued Functions

- An extended real-valued function is called convex if $\mathrm{epi}(f)$ is convex.
- $f : \mathbb{E} \to (-\infty, \infty]$ is convex $\Leftrightarrow$ $\mathrm{dom}(f)$ is convex and the real-valued function $\tilde{f} : \mathrm{dom}(f) \to \mathbb{R}$ which is the restriction of f to $\mathrm{dom}(f)$ is convex over $\mathrm{dom}(f)$.
- Result: A proper function $f : \mathbb{E} \to (-\infty, \infty]$ is convex iff

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \text{ for all } \lambda \in [0,1], \mathbf{x}, \mathbf{y} \in \mathbb{E}$$

- Jensen's inequality

$$f\left(\sum_{i=1}^{k} \lambda_i \mathbf{x}_i\right) \leq \sum_{i=1}^{k} \lambda_i f_i(\mathbf{x}_i)$$

for any $\boldsymbol{\lambda} \in \Delta_k$ (k being an arbitrary positive integer), $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k \in \mathbb{E}$.

Operations Preserving Convexity

Theorem.

(a) Let $\mathcal{A} : \mathbb{E} \to \mathbb{V}$ be a linear transformation from $\mathbb{E}$ to $\mathbb{V}$ and $\mathbf{b} \in \mathbb{V}$, and let $f : \mathbb{V} \to (-\infty, \infty]$ be convex. Then $g : \mathbb{E} \to (-\infty, \infty]$ given by

$$g(\mathbf{x}) = f(\mathcal{A}(\mathbf{x}) + \mathbf{b})$$

is convex.

(b) Let $f_1, f_2, \ldots, f_m : \mathbb{E} \to (-\infty, \infty]$ be convex, and let $\alpha_1, \alpha_2, \ldots, \alpha_m \in \mathbb{R}_+$. Then the function $\sum_{i=1}^m \alpha_i f_i$ is convex.

(c) Let $f_i : \mathbb{E} \to (-\infty, \infty], i \in I$ be convex, where I is a given index set. Then the function

$$f(\mathbf{x}) = \max_{i \in I} f_i(\mathbf{x})$$

is convex.

Closedness Vs. Continuity

Closed functions are not necessarily continuous, but...

- If $f : \mathbb{E} \to [-\infty, \infty]$ is continuous over $\text{dom}(f)$, which is assumed to be closed, then it is closed.
- 1D closed and convex functions are always continuous over their domain.
- Not correct for multi-dimensional functions...

Example: the l_0-norm function $f : \mathbb{R}^n \to \mathbb{R}$ given by

$$f(\mathbf{x}) = \|\mathbf{x}\|_0 \equiv \#\{i : x_i \neq 0\}.$$

f is closed but not continuous.

Support Functions

- Let $C \subseteq \mathbb{E}$ be nonempty. Then the support function of C, $\sigma_C : \mathbb{E} \to (-\infty, \infty]$ is given by

$$\sigma_C(\mathbf{y}) \equiv \max_{\mathbf{y} \in C} \langle \mathbf{y}, \mathbf{x} \rangle.$$

Theorem. Let $C \subseteq \mathbb{E}$ be a nonempty set. Then σ_C is a closed and convex function.

Proof. σ_C is a maximum of convex functions.

Examples of Support Functions

C	$\sigma_C(\mathbf{y})$	assumptions	Example No.
$\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$	$\max_{i=1,2,\ldots,n} \langle \mathbf{b}_i, \mathbf{y} \rangle$	$\mathbf{b}_i \in \mathbb{E}$	1
K	$\delta_{K^\circ}(\mathbf{y})$	K – cone	2
$\mathbb{R}^n_+$	$\delta_{\mathbb{R}^n_-}(\mathbf{y})$	$\mathbb{E} = \mathbb{R}^n$	3
Δ_n	$\max\{y_1, y_2, \ldots, y_n\}$	$\mathbb{E} = \mathbb{R}^n$	4
$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{0}\}$	$\delta_{\{\mathbf{A}^T \boldsymbol{\lambda} : \boldsymbol{\lambda} \in \mathbb{R}^m_+\}}(\mathbf{y})$	$\mathbb{E} = \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{m \times n}$	5
$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{B}\mathbf{x} = \mathbf{b}\}$	$\langle \mathbf{y}, \mathbf{x}_0 \rangle + \delta_{\text{Range}(\mathbf{B}^T)}(\mathbf{y})$	$\mathbb{E} = \mathbb{R}^n$, $\mathbf{B} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{B}\mathbf{x}_0 = \mathbf{b}$	6
$B_{\|\cdot\|}[\mathbf{0}, 1]$	$\|\mathbf{y}\|_*$	$\|\cdot\|$ - arbitrary norm	7

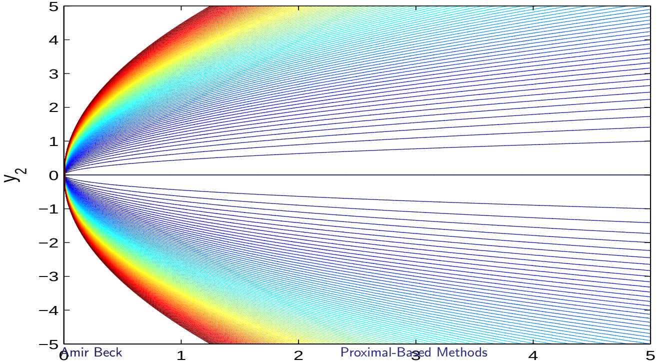
A Discontinuous Closed and Convex Function

If

$$C = \left\{ (x_1, x_2) : x_1 + \frac{x_2^2}{2} \le 0 \right\}.$$

Then

$$\sigma_C(\mathbf{y}) = \begin{cases} \frac{y_2^2}{2y_1}, & y_1 > 0 \\ 0, & y_1 = y_2 = 0 \\ \infty, & \text{else.} \end{cases}$$



Subgradients

- D. P. Bertsekas, A. Nedic and A. E. Ozdaglar, *Convex analysis and optimization* (2013).
- J. M. Borwein and A. S. Lewis, *Convex analysis and nonlinear optimization* (2006).
- J. B. Hiriart-Urruty and C. Lemarechal. *Convex analysis and minimization algorithms. I* (1996).
- Y. Nesterov. *Introductory lectures on convex optimization* (2004).
- R. T. Rockafellar, *Convex analysis* (1970).

Subgradients

- Definition: Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper function, and let $\mathbf{x} \in \text{dom}(f)$. A vector $\mathbf{g} \in \mathbb{E}$ is called a subgradient of f at $\mathbf{x}$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \text{ for all } \mathbf{y} \in \mathbb{E}.$$

- The set of all subgradients of f at $\mathbf{x}$ is called the subdifferential of f at $\mathbf{x}$ and is denoted by $\partial f(\mathbf{x})$:

$$\partial f(\mathbf{x}) \equiv \{\mathbf{g} \in \mathbb{E} : f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \text{ for all } \mathbf{y} \in \mathbb{E}\}.$$

When $\mathbf{x} \notin \text{dom}(f)$, we define $\partial f(\mathbf{x}) = \emptyset$.

Closedness and Convexity of the Subdifferential Set

Theorem. Let $f : \mathbb{E} \to (\infty, \infty]$ be an extended real-valued function. Then the set $\partial f(\mathbf{x})$ is closed and convex for any $\mathbf{x} \in \mathbb{E}$.

Proof. For any $\mathbf{x} \in \mathbb{E}$,

$$\partial f(\mathbf{x}) = \bigcap_{\mathbf{y} \in \mathbb{E}} H_{\mathbf{y}},$$

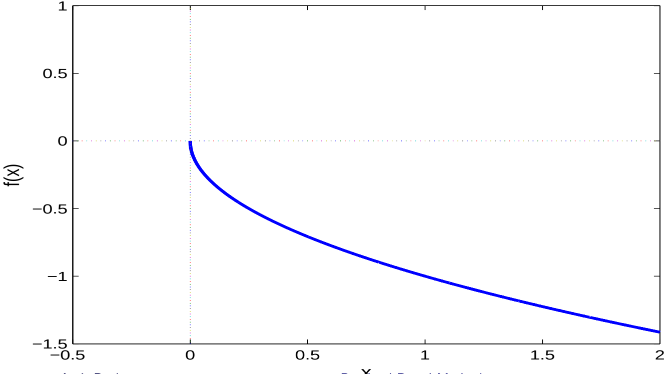
where $H_{\mathbf{y}} = \{\mathbf{g} \in \mathbb{E} : f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle\}$. Since the sets $H_{\mathbf{y}}$ are half-spaces, and in particular, closed and convex, it follows that $\partial f(\mathbf{x})$ is closed and convex. $\square$

Subdifferentiability

- If $\partial f(\mathbf{x}) \neq \emptyset$, f it is called subdifferentiable at $\mathbf{x}$.
- $$\text{dom}(\partial f) \equiv \{\mathbf{x} \in \mathbb{E} : \partial f(\mathbf{x}) \neq \emptyset\}.$$

Example:

$$f(x) = \begin{cases} -\sqrt{x}, & x \geq 0, \\ \infty, & \text{else.} \end{cases}$$



Existence and Boundedness of $\partial f(\mathbf{x})$

Theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function.

- If $\tilde{\mathbf{x}} \in \text{int}(\text{dom}(f))$, then $\partial f(\tilde{\mathbf{x}})$ is nonempty and bounded.
- If $\tilde{\mathbf{x}} \in \text{ri}(\text{dom}(f))$, then $\partial f(\tilde{\mathbf{x}})$ is nonempty.

Corollary. Let $f : \mathbb{E} \to \mathbb{R}$ be a convex function. Then f is subdifferentiable over $\mathbb{E}$.

Theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function, and assume that $X \subseteq \text{int}(\text{dom}(f))$ is nonempty and compact. Then $Y = \bigcup_{\mathbf{x} \in X} \partial f(\mathbf{x})$ is nonempty and bounded.

The Directional Derivative

- Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper extended real-valued function and let $\mathbf{x} \in \text{int}(\text{dom}(f))$. Suppose that $\mathbf{0} \neq \mathbf{d} \in \mathbb{E}$. The directional derivative at $\mathbf{x}$ in the direction $\mathbf{0} \neq \mathbf{d} \in \mathbb{E}$, if exists, is defined by

$$f'(\mathbf{x}; \mathbf{d}) = \lim_{\alpha \to 0^+} \frac{f(\mathbf{x} + \alpha \mathbf{d}) - f(\mathbf{x})}{\alpha}.$$

Theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function, and let $\mathbf{x} \in \text{int}(\text{dom}(f))$. Then for any $\mathbf{d} \in \mathbb{E}$, the directional derivative $f'(\mathbf{x}; \mathbf{d})$ exists.

Differentiability

Definition. For a given function $f : \mathbb{E} \to (-\infty, \infty]$, and $\mathbf{x} \in \text{int}(\text{dom}(f))$, we say that f is differentiable at $\mathbf{x}$ if there exists $\mathbf{g} \in \mathbb{E}$ such that

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{h} \rangle + o(\|\mathbf{h}\|).$$

In other words, $\lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - \langle \mathbf{g}, \mathbf{h} \rangle}{\|\mathbf{h}\|} = 0.$

$\mathbf{g}$ is called the gradient, and is denoted by $\nabla f(\mathbf{x})$

Theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$, and suppose that f is differentiable at $\mathbf{x} \in \text{int}(\text{dom}\, f)$. Then for any $\mathbf{d} \neq \mathbf{0}$

$$f'(\mathbf{x}; \mathbf{d}) = \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle.$$

Proof. $0 = \lim_{\alpha \to 0^+} \frac{f(\mathbf{x}+\alpha\mathbf{d}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \alpha\mathbf{d} \rangle}{\|\alpha\mathbf{d}\|} = \frac{f'(\mathbf{x};\mathbf{d}) - \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle}{\|\mathbf{d}\|}$, and hence $f'(\mathbf{x}; \mathbf{d}) = \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle$. $\square$

The Subdifferential at Differentiability Points

Theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function, and let $\mathbf{x} \in \text{int}(\text{dom}(f))$. If f is differentiable at $\mathbf{x}$, then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$. Conversely, if f has a unique subgradient at $\mathbf{x}$, then f is differentiable at $\mathbf{x}$ and $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.

Example: $f(\mathbf{x}) = \|\mathbf{x}\|_2$ ($\mathbb{E} = \mathbb{R}^n$). Then $\partial f(\mathbf{x}) = \begin{cases} \left\{ \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \right\}, & \mathbf{x} \neq \mathbf{0}, \\ B_{\|\cdot\|_2}[\mathbf{0}, 1], & \mathbf{x} = \mathbf{0}. \end{cases}$

What is the Gradient?

- **Example 1:** $\mathbb{E} = \mathbb{R}^n$ with $\langle \mathbf{x}, \mathbf{y} \rangle \equiv \mathbf{x}^T\mathbf{y}$: $\nabla f(\mathbf{x}) = D_f(\mathbf{x})$

$$D_f(\mathbf{x}) \equiv \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \frac{\partial f}{\partial x_2}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{pmatrix}.$$

- **Example 2:** $\mathbb{E} = \mathbb{R}^n$ with $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T\mathbf{H}\mathbf{y}$ with $\mathbf{H} \in \mathbb{S}^n_{++}$: $\nabla f(\mathbf{x}) = \mathbf{H}^{-1}D_f(\mathbf{x})$.

Subdifferential Calculus

Theorem. Let $f_1, f_2 : \mathbb{R}^n \to (-\infty, \infty]$ be proper extended real-valued convex functions. Let $\mathbf{x} \in \mathrm{dom}(f_1) \cap \mathrm{dom}(f_2)$. Then

(a) The following inclusion holds (**weak result**):

$$\partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}) \subseteq \partial(f_1 + f_2)(\mathbf{x})$$

(b) If in addition either $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f_1)) \cap \mathrm{int}(\mathrm{dom}(f_2))$, then (**strong result**):

$$\partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}) = \partial(f_1 + f_2)(\mathbf{x}).$$

Sum Rule of Subdifferential Calculus - General Result

Theorem. Let $f_1, f_2, \ldots, f_m$ be proper convex functions and assume that $\bigcap_{i=1}^m \text{ri}(\text{dom}\, f_i) \neq \emptyset$. Then for any $\mathbf{x}$

$$\partial f(\mathbf{x}) = \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}) + \ldots + f_m(\mathbf{x})$$

Subdifferential Calculus - Affine Change of Variables

Theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function and $\mathcal{A} : \mathbb{V} \to \mathbb{E}$ be a linear transformation. Let $h(\mathbf{x}) = f(\mathcal{A}(\mathbf{x}) + \mathbf{b})$ with $\mathbf{b} \in \mathbb{E}$. Assume that h is proper:

$$\mathrm{dom}(h) = \{\mathbf{x} \in \mathbb{V} : \mathcal{A}(\mathbf{x}) + \mathbf{b} \in \mathrm{dom}(f)\} \neq \emptyset.$$

(a) **(weak affine transformation rule of subdifferential calculus)** For any $\mathbf{x} \in \mathrm{dom}(h)$,

$$\mathcal{A}^T(\partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b})) \subseteq \partial h(\mathbf{x}).$$

(b) **(affine transformation rule of subdifferential calculus)** If $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(h))$ and $\mathcal{A}(\mathbf{x}) + \mathbf{b} \in \mathrm{int}(\mathrm{dom}(f))$, then

$$\partial h(\mathbf{x}) = \mathcal{A}^T(\partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b})).$$

Chain Rule of Subdifferential Calculus

Theorem Let $f : \mathbb{E} \to \mathbb{R}$ be a convex function and let $g : \mathbb{R} \to \mathbb{R}$ be a nondecreasing convex function. Let $\mathbf{x} \in \mathbb{E}$ and suppose that g is differentiable at the point $f(\mathbf{x})$. Let $h = g \circ f$. Then

$$\partial h(\mathbf{x}) = g'(f(\mathbf{x}))\partial f(\mathbf{x}).$$

Max Rule of Subdifferential Calculus

Lemma. Let $f_1, f_2, \ldots, f_m : \mathbb{E} \to (-\infty, \infty]$ be proper extended real-valued convex functions and let

$$f(\mathbf{x}) \equiv \max\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\}.$$

Let $\mathbf{x} \in \bigcap_{i=1}^m \operatorname{int}(\mathrm{dom}(f_i))$. Then

$$\partial f(\mathbf{x}) = \mathrm{conv}\left(\bigcup_{i \in I(\mathbf{x})} \partial f_i(\mathbf{x})\right),$$

where

$$I(\mathbf{x}) = \{i \in \{1, 2, \ldots, m\} : f_i(\mathbf{x}) = f(\mathbf{x})\}.$$

Lipschitz Continuity and Boundedness of Subgradients

Theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper and convex function. Suppose that $X \subseteq \mathrm{int}(\mathrm{dom}\, f)$. Consider the following two claims:

(i) $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in X$;

(ii) $\|\mathbf{g}\|_* \leq L$ for any $\mathbf{g} \in \partial f(\mathbf{x}), \mathbf{x} \in X$.

Then

(a) the implication $(ii) \Rightarrow (i)$ holds;

(b) if X is open then (i) holds if and only if (ii) holds.

Fermat's Optimality Condition

Theorem. Let $f : \mathbb{E} \rightarrow (-\infty, \infty]$ be an extended real-valued convex function. Then

$$\mathbf{x}^* \in \operatorname{argmin}\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\} \tag{1}$$

if and only if

$$\mathbf{0} \in \partial f(\mathbf{x}^*)$$

Proof. $\mathbf{0} \in \partial f(\mathbf{x}^*)$ is satisfied iff

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \langle \mathbf{0}, \mathbf{x} - \mathbf{x}^* \rangle \text{ for any } \mathbf{x} \in \operatorname{dom}(f),$$

which is the the same as (1).

Fermat-Weber Problem

Given m different points in $\mathbb{R}^d$, $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\}$ ("anchors") and m positive weights $\omega_1, \omega_2, \ldots, \omega_m$, the Fermat-Weber problem is given by

$$\text{(FW)} \quad \min_{\mathbf{x} \in \mathbb{R}^d} \left\{ f(\mathbf{x}) \equiv \sum_{i=1}^m \omega_i \|\mathbf{x} - \mathbf{a}_i\|_2 \right\}.$$

- $$\partial f(\mathbf{x}) = \sum_{i=1}^m \partial f_i(\mathbf{x}) = \begin{cases} \sum_{i=1}^m \omega_i \frac{\mathbf{x}-\mathbf{a}_i}{\|\mathbf{x}-\mathbf{a}_i\|_2}, & \mathbf{x} \notin \mathcal{A}, \\ \sum_{i=1, i\neq j}^m \omega_i \frac{\mathbf{x}-\mathbf{a}_i}{\|\mathbf{x}-\mathbf{a}_i\|_2} + B[\mathbf{0}, \omega_j], & \mathbf{x} = \mathbf{a}_j (j \in [m]). \end{cases}$$
- By Fermat's optimality optimality condition, $\mathbf{x}^*$ is an optimal solution iff $\mathbf{0} \in \partial f(\mathbf{x}^*)$, meaning iff
- $\mathbf{x}^* \notin \mathcal{A}$ and $\sum_{i=1}^m \omega_i \frac{\mathbf{x}^*-\mathbf{a}_i}{\|\mathbf{x}^*-\mathbf{a}_i\|_2} = \mathbf{0}$ or for some $j \in \{1, 2, \ldots, m\}$ $\mathbf{x}^* = \mathbf{a}_j$ and $\left\| \sum_{i=1, i\neq j}^m \omega_i \frac{\mathbf{x}^*-\mathbf{a}_i}{\|\mathbf{x}^*-\mathbf{a}_i\|_2} \right\|_2 \le \omega_j$.

[Sturm, 1884] [Weiszfeld, 1937]

Optimality Conditions for the Composite Model (Mixed Convex/Nonconvex)

Theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be proper, and let $g : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function such that $\text{dom}(g) \subseteq \text{int}(\text{dom}(f))$. Consider the problem

$$\text{(P)} \quad \min f(\mathbf{x}) + g(\mathbf{x}).$$

(a) **(necessary condition)** If $\mathbf{x}^* \in \text{dom}(g)$ is a local optimal solution of (P), and f is differentiable at $\mathbf{x}^*$, then

$$-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*). \tag{2}$$

(b) **(necessary and sufficient condition for convex problems)** Suppose that f is convex. If f is differentiable at $\mathbf{x}^* \in \text{dom}(g)$, then $\mathbf{x}^*$ is a global optimal solution of (P) if and only if (2) is satisfied.

Stationarity in Composite Models

$$\text{(P)} \quad \min f(\mathbf{x}) + g(\mathbf{x}).$$

- $f : \mathbb{E} \to (-\infty, \infty]$ proper.
- $g : \mathbb{E} \to (-\infty, \infty]$ proper convex.
- $\text{dom}(g) \subseteq \text{int}(\text{dom}(f))$.

Definition A point $\mathbf{x}^* \in \text{dom}\, g$ in which f is differentiable is called a stationarity point of (P) if $-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*)$

Example: If $g(\mathbf{x}) = \delta_C(\mathbf{x})$ for convex C, then stationarity is the same as

$$\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0$$

Example: $\min f(\mathbf{x}) + \lambda \|\mathbf{x}\|_1$ ($f : \mathbb{R}^n \to \mathbb{R}$), then stationarity is

$$\frac{\partial f(\mathbf{x}^*)}{\partial x_i} \begin{cases} = -\lambda, & x_i^* > 0, \\ = \lambda, & x_i^* < 0, \\ \in [-\lambda, \lambda], & x_i^* = 0. \end{cases}$$

Conjugate Functions

- D. P. Bertsekas, A. Nedic and A. E. Ozdaglar, *Convex analysis and optimization* (2013).
- J. M. Borwein and A. S. Lewis, *Convex analysis and nonlinear optimization* (2006).
- J. B. Hiriart-Urruty and C. Lemarechal. *Convex analysis and minimization algorithms. I* (1996).
- R. T. Rockafellar, *Convex analysis* (1970).

Conjugate Functions

Definition. Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper extended real-valued function. The function $f : \mathbb{E} \to [-\infty, \infty]$ defined by

$$f^*(\mathbf{y}) = \max_{\mathbf{x}\in\mathbb{E}}\{\langle \mathbf{y}, \mathbf{x}\rangle - f(\mathbf{x})\}.$$

is called the conjugate function of f.

Result: Conjugate functions are **always** closed and convex (regardless of the properties of f)

Example: $f = \delta_C$, where $C \subseteq \mathbb{E}$ is nonempty. Then for any $\mathbf{y} \in \mathbb{E}$

$$f^*(\mathbf{y}) = \max_{\mathbf{x}\in\mathbb{E}}\{\langle \mathbf{y}, \mathbf{x}\rangle - \delta_C(\mathbf{x})\} = \max_{\mathbf{x}\in C}\langle \mathbf{y}, \mathbf{x}\rangle = \sigma_C(\mathbf{y}).$$

$$\boxed{\delta_C^* = \sigma_C.}$$

The Biconjugate

The conjugacy operation can be invoked twice resulting with the biconjugacy operation. Specifically, for a function f we define

$$f^{**}(\mathbf{x}) = \max_{\mathbf{y} \in \mathbb{E}} \langle \mathbf{x}, \mathbf{y} \rangle - f^*(\mathbf{y})$$

Theorem ($f \geq f^{**}$). Let $f : \mathbb{E} \to [-\infty, \infty]$ be an extended real-valued function. Then $f(\mathbf{x}) \geq f^{**}(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{E}$.

Theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be a closed and proper extended real-valued function. Then $f^{**} = f$.

Fenchel's Inequality

Theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be an extended real-valued proper function. Then for any $\mathbf{x} \in \mathbb{E}, \mathbf{y} \in \mathbb{E}$

$$f(\mathbf{x}) + f^*(\mathbf{y}) \geq \langle \mathbf{y}, \mathbf{x} \rangle.$$

Simple Calculus Rules

function definition	conjugate
$g(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \sum_{i=1}^m f_i(\mathbf{x}_i)$	$g^*(\mathbf{y}_1, \ldots, \mathbf{y}_m) = \sum_{i=1}^m f_i^*(\mathbf{y}_i)$
$g(\mathbf{x}) = \alpha f(\mathbf{x})$	$g^*(\mathbf{y}) = \alpha f^*(\mathbf{y}/\alpha)$
$g(\mathbf{x}) = \alpha f(\mathbf{x}/\alpha)$	$g^*(\mathbf{y}) = \alpha f^*(\mathbf{y})$
$f(\mathcal{A}(\mathbf{x}-\mathbf{a})) + \langle \mathbf{b}, \mathbf{x} \rangle + c$	$f^*\left((\mathcal{A}^T)^{-1}(\mathbf{y}-\mathbf{b})\right) + \langle \mathbf{a}, \mathbf{y} \rangle - c - \langle \mathbf{a}, \mathbf{b} \rangle$

Conjugates of Simple Functions

function (f)	dom f	conjugate (f^*)	assumptions
$\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}+\mathbf{b}^T\mathbf{x}+c$	$\mathbb{R}^n$	$\frac{1}{2}(\mathbf{y}-\mathbf{b})^T\mathbf{A}^{-1}(\mathbf{y}-\mathbf{b})-c$	$\mathbf{A}\succ\mathbf{0}, \mathbf{A}\in\mathbb{R}^{n\times n}$, $\mathbf{b}\in\mathbb{R}^n$, $c\in\mathbb{R}$
$\sum_{i=1}^n x_i\log x_i$	$\mathbb{R}^n_+$	$\sum_{i=1}^n e^{y_i-1}$	–
$\sum_{i=1}^n x_i\log x_i$	Δ_n	$\log\left(\sum_{i=1}^n e^{y_i}\right)$	–
$\log\left(\sum_{i=1}^n e^{x_i}\right)$	$\mathbb{R}^n$	$\sum_{i=1}^n y_i\log y_i$ (dom $f^*=\Delta_n$)	–
$\delta_C(\mathbf{x})$	C	$\sigma_C(\mathbf{x})$	$\emptyset\neq C$ arbitrary
$\sigma_C(\mathbf{x})$	$\mathbb{R}^n$	$\delta_C(\mathbf{x})$	$\emptyset\neq C$ closed, convex
$\lVert\mathbf{x}\rVert$	$\mathbb{R}^n$	$\delta_{B_{\lVert\cdot\rVert_*}[\mathbf{0},1]}$	$\lVert\cdot\rVert$ arbitrary norm
$-\sqrt{1-\lVert\mathbf{x}\rVert^2}$	$B_{\lVert\cdot\rVert}[\mathbf{0},1]$	$\sqrt{\lVert\mathbf{y}\rVert_*^2+1}$	$\lVert\cdot\rVert$ arbitrary norm
$\frac{1}{p}\lvert x\rvert^p$	$\mathbb{R}$	$\frac{1}{q}\lvert y\rvert^q$	$p>1, \frac{1}{p}+\frac{1}{q}=1$
$\frac{1}{2}\lVert\mathbf{x}\rVert^2$	$\mathbb{R}^n$	$\frac{1}{2}\lVert\mathbf{y}\rVert_*^2$	$\lVert\cdot\rVert$ arbitrary norm

Conjugate Subgradient Theorem

Theorem. Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a proper convex extended real-valued function. The following two claims are equivalent for any $\mathbf{x} \in \mathbb{E}, \mathbf{y} \in \mathbb{E}$:

(i) $\langle \mathbf{x}, \mathbf{y} \rangle = f(\mathbf{x}) + f^*(\mathbf{y})$.

(ii) $\mathbf{y} \in \partial f(\mathbf{x})$.

If, in addition f is closed, then (i) and (ii) are equivalent to

(iii) $\mathbf{x} \in \partial f^*(\mathbf{y})$.

- If f is proper closed and convex, the conjugate subgradient theorem can be written as

$$\begin{aligned} \partial f^*(\mathbf{y}) &= \underset{\mathbf{x}}{\operatorname{argmax}} \left\{ \langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x}) \right\}, \\ \partial f(\mathbf{x}) &= \underset{\mathbf{y}}{\operatorname{argmax}} \left\{ \langle \mathbf{x}, \mathbf{y} \rangle - f^*(\mathbf{y}) \right\} \end{aligned}$$

Fenchel's Duality Theorem

$$\text{(P)} \quad \min_{\mathbf{x}\in\mathbb{E}} f(\mathbf{x}) + g(\mathbf{x}).$$

Lagrangian duality:

- $\min_{\mathbf{x},\mathbf{z}\in\mathbb{E}}\{f(\mathbf{x}) + g(\mathbf{z}) : \mathbf{x} = \mathbf{z}\}$
- Lagrangian:
  $L(\mathbf{x}, \mathbf{z}; \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{z}) + \langle \mathbf{y}, \mathbf{z} - \mathbf{x}\rangle = -\left[\langle \mathbf{y}, \mathbf{x}\rangle - f(\mathbf{x})\right] - \left[\langle -\mathbf{y}, \mathbf{z}\rangle - g(\mathbf{z})\right].$
- Dual objective function: $q(\mathbf{y}) = \min_{\mathbf{x},\mathbf{z}} L(\mathbf{x}, \mathbf{z}; \mathbf{y}) = -f^*(\mathbf{y}) - g^*(-\mathbf{y})$

Fenchel's dual problem:

$$\text{(D)} \quad \max_{\mathbf{y}\in\mathbb{E}^*}\{-f^*(\mathbf{y}) - g^*(-\mathbf{y})\}.$$

Theorem (Fenchel's duality theorem) Let $f, g : \mathbb{E} \to (-\infty, \infty]$ be proper convex functions. If $\text{ri}(\text{dom}(f)) \cap \text{ri}(\text{dom}(g)) \neq \emptyset$, then

$$\min_{\mathbf{x}\in\mathbb{E}}\{f(\mathbf{x}) + g(\mathbf{x})\} = \max_{\mathbf{y}\in\mathbb{E}^*}\{-f^*(\mathbf{y}) - g^*(-\mathbf{y})\},$$

and the maximum in the right-hand problem is attained whenever it is finite.

The Proximal Operator

- J. J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France (1965).
- H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces* (2011).
- P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward backward splitting*, Multiscale Model. Simul. (2005).
- N. Parikh and S. Boyd, *Proximal algorithms*, Foundations and Trends in Optimization (2014).

The Proximal Operator

Definition. Given a closed, proper and convex function g, the proximal mapping of g is defined by

$$\mathrm{prox}_g(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{u} \in \mathbb{E}} \left\{ g(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2 \right\}.$$

Examples

- **Constant.** If $f \equiv c$ for some $c \in \mathbb{R}$, then

$$\mathrm{prox}_f(\mathbf{x}) = \underset{\mathbf{u}\in\mathbb{E}}{\operatorname{argmin}} \left\{ c + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2 \right\} = \mathbf{x}$$

  The identity mapping.

- **Affine.** Let $f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + b$, where $\mathbf{a} \in \mathbb{E}$ and $b \in \mathbb{R}$. Then

$$\begin{aligned} \mathrm{prox}_f(\mathbf{x}) &= \underset{\mathbf{u}\in\mathbb{E}}{\operatorname{argmin}} \left\{ \langle \mathbf{a}, \mathbf{u} \rangle + b + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2 \right\} \\ &= \mathbf{x} - \mathbf{a}. \end{aligned}$$

- Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$, where $\mathbf{A} \in \mathbb{S}^n_+, \mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. The vector $\mathrm{prox}_f(\mathbf{x})$ is the solution of

$$\min_{\mathbf{u}\in\mathbb{E}} \left\{ \frac{1}{2}\mathbf{u}^T\mathbf{A}\mathbf{u} + \mathbf{b}^T\mathbf{u} + c + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2 \right\}.$$

  The optimal solution is attained at $\mathbf{u}$ satisfying $(\mathbf{A} + \mathbf{I})\mathbf{u} = \mathbf{x} - \mathbf{b}$, and hence

$$\mathrm{prox}_f(\mathbf{x}) = \mathbf{u} = (\mathbf{A} + \mathbf{I})^{-1}(\mathbf{x} - \mathbf{b}).$$

The Orthogonal Projection

- Definition. Given a nonempty closed and convex set $C \subseteq \mathbb{E}$ and $\mathbf{x} \in \mathbb{E}$, the orthogonal projection operator $P_C : \mathbb{E} \to C$ is defined by

$$P_C(\mathbf{x}) \equiv \underset{\mathbf{y} \in C}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{x}\|.$$

First projection theorem. Let $C \subseteq \mathbb{E}$ be a nonempty closed convex set. Then $P_C(\mathbf{x})$ is a singleton.

Second projection theorem. Let $C \subseteq \mathbb{E}$ be a nonempty closed and convex set. Let $\mathbf{u} \in C$. Then $\mathbf{u} = P_C(\mathbf{x})$ if and only if

$$\langle \mathbf{x} - \mathbf{u}, \mathbf{y} - \mathbf{u} \rangle \leq 0 \text{ for any } \mathbf{y} \in C.$$

Prox of Indicator = Orthogonal Projection

- If $C \subseteq \mathbb{E}$ is nonempty, then $\mathrm{prox}_{\delta_C} = P_C$

$$\mathrm{prox}_{\delta_C}(\mathbf{x}) = \underset{\mathbf{u}\in\mathbb{E}}{\mathrm{argmin}} \left\{ \delta_C(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2 \right\} = \underset{\mathbf{u}\in C}{\mathrm{argmin}} \|\mathbf{u}-\mathbf{x}\|^2 = P_C(\mathbf{x}).$$

First prox theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper closed and convex function. Then $\mathrm{prox}_f(\mathbf{x})$ is a singleton for any $\mathbf{x} \in \mathbb{E}$.

Proof?

Strongly Convex Functions

Definition. A function $f : \mathbb{E} \to (-\infty, \infty]$ is called σ-strongly convex for a given $\sigma > 0$, if $\mathrm{dom}(f)$ is convex and the following inequality holds for any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$ and $\lambda \in [0,1]$:

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}) - \frac{1}{2}\sigma\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2.$$

- A function is strongly convex if it is σ-strongly convex for some $\sigma > 0$.

Theorem. $f : \mathbb{E} \to (-\infty, \infty]$ is a strongly convex function if and only if the function $f(\cdot) - \frac{\sigma}{2}\|\cdot\|^2$ is convex.

- The proof is extremely straightforward.
- The above characterization is relevant only for Euclidean spaces.
- σ-strongly convex+convex is σ-strongly convex.

Example: $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$ $(\mathbf{A} \in \mathbb{S}^n, \mathbf{b} \in \mathbb{R}^n, c \in \mathbb{R})$ is strongly convex with parameter $\lambda_{\min}(\mathbf{A})$.

First Order Characterizations of Strong Convexity

Theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper closed and convex function. Then for a given $\sigma > 0$, the following three claims are equivalent:

(i) f is σ-strongly convex.

(ii)
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle + \frac{\sigma}{2}\|\mathbf{y} - \mathbf{x}\|^2$$
for any $\mathbf{x} \in \operatorname{dom}(\partial f), \mathbf{y} \in \operatorname{dom}(f)$ and $\mathbf{g} \in \partial f(\mathbf{x})$.

(iii)
$$\langle \mathbf{g}_{\mathbf{x}} - \mathbf{g}_{\mathbf{y}}, \mathbf{x} - \mathbf{y} \rangle \geq \sigma \|\mathbf{x} - \mathbf{y}\|^2$$
for any $\mathbf{x}, \mathbf{y} \in \operatorname{dom}(\partial f)$ and $\mathbf{g}_{\mathbf{x}} \in \partial f(\mathbf{x}), \mathbf{g}_{\mathbf{y}} \in \partial f(\mathbf{y})$.

Existence and Uniqueness of a Minimizer of Closed Strongly Convex Functions

Theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper closed and σ-strongly convex function ($\sigma > 0$). Then

(a) f has a unique minimizer.

(b) $f(\mathbf{x}) - f(\mathbf{x}^*) \geq \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^*\|^2$ for all $\mathbf{x} \in \text{dom}(f)$, where $\mathbf{x}^*$ is the unique minimizer of f.

Conclusion: the first prox theorem.

First prox theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper closed and convex function. Then $\operatorname{prox}_f(\mathbf{x})$ is a singleton for any $\mathbf{x} \in \mathbb{E}$.

Proof.

- For any $\mathbf{x} \in \mathbb{E}$,
$$\operatorname{prox}_f(\mathbf{x}) = \underset{\mathbf{u} \in \mathbb{E}}{\text{argmin}}\, \tilde{f}(\mathbf{u}, \mathbf{x}), \tag{3}$$
where $\tilde{f}(\mathbf{u}, \mathbf{x}) = f(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2$.
- $\tilde{f}(\cdot, \mathbf{x})$ is a proper closed and 1-strongly convex function.
- Therefore, there exists a unique minimizer to the problem in (3).

Necessity of the Conditions in the First Prox Theorem

• When f is not convex and/or closed, the prox is not guaranteed to uniquely exist, or even to exist at all.

$$
\begin{aligned}
g_1(x) &\equiv 0, \\
g_2(x) &= \begin{cases} 0, & x \neq 0, \\ -\lambda, & x = 0, \end{cases} \\
g_3(x) &= \begin{cases} 0, & x \neq 0, \\ \lambda, & x = 0. \end{cases}
\end{aligned}
$$

$$
\mathrm{prox}_{g_1}(x) = x, \mathrm{prox}_{g_2}(x) = \begin{cases} \{0\}, & |x| < \sqrt{2\lambda}, \\ \{x\}, & |x| > \sqrt{2\lambda}, \\ \{0, x\}, & |x| = \sqrt{2\lambda}. \end{cases}, \mathrm{prox}_{g_3}(x) = \begin{cases} \{x\}, & x \neq 0, \\ \emptyset, & x = 0. \end{cases}
$$

- Uniquness is not guaranteed in any case.
- Existence is guaranteed whenever f is proper closed and the function $\mathbf{u} \mapsto f(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2$ is coercive.

Basic Calculus Rules

$f(\mathbf{x})$	$\mathrm{prox}_f(\mathbf{x})$	assumptions
$\sum_{i=1}^m f_i(\mathbf{x}_i)$	$\mathrm{prox}_{f_1}(\mathbf{x}_1) \times \cdots \times \mathrm{prox}_{f_m}(\mathbf{x}_m)$	
$g(\lambda\mathbf{x} + \mathbf{a})$	$\frac{1}{\lambda}\left[\mathrm{prox}_{\lambda^2 g}(\mathbf{a} + \lambda\mathbf{x}) - \mathbf{a}\right]$	$\lambda \neq 0, \mathbf{a} \in \mathbb{E}$, g proper
$\lambda g(\mathbf{x}/\lambda)$	$\lambda \mathrm{prox}_{g/\lambda}(\mathbf{x}/\lambda)$	$\lambda > 0$, g proper
$g(\mathbf{x}) + \frac{c}{2}\|\mathbf{x}\|^2 + \langle \mathbf{a}, \mathbf{x}\rangle + \gamma$	$\mathrm{prox}_{\frac{1}{c+1}g}\left(\frac{\mathbf{x}-\mathbf{a}}{c+1}\right)$	$\mathbf{a} \in \mathbb{E}, c > 0, \gamma \in \mathbb{R}$, g proper
$g(\mathcal{A}(\mathbf{x}) + \mathbf{b})$	$\mathbf{x} + \frac{1}{\alpha}\mathcal{A}^T(\mathrm{prox}_{\alpha g}(\mathcal{A}(\mathbf{x}) + \mathbf{b}) - \mathcal{A}(\mathbf{x}) - \mathbf{b})$	$\mathbf{b} \in \mathbb{R}^m$, $\mathcal{A} : \mathbb{V} \to \mathbb{R}^m$, g closed proper convex, $\mathcal{A} \circ \mathcal{A}^T = \alpha I$, $\alpha > 0$
$g(\|\mathbf{x}\|)$	$\mathrm{prox}_g(\|\mathbf{x}\|)\frac{\mathbf{x}}{\|\mathbf{x}\|}$, $\mathbf{x} \neq \mathbf{0}$ $\{\mathbf{u} : \|\mathbf{u}\| = \mathrm{prox}_g(0)\}$, $\mathbf{x} = \mathbf{0}$	g proper closed convex, $\mathrm{dom}(g) \subseteq [0, \infty)$

Examples or Prox Computations

f	dom f	prox_f	assumptions
$\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}+\mathbf{b}^T\mathbf{x}+c$	$\mathbb{R}^n$	$(\mathbf{A}+\mathbf{I})^{-1}(\mathbf{x}-\mathbf{b})$	$\mathbf{A}\in\mathbb{S}^n_{++}$, $\mathbf{b}\in\mathbb{R}^n$, $c\in\mathbb{R}$
$\lambda\|\mathbf{x}\|$	$\mathbb{E}$	$\left[1-\frac{\lambda}{\|\mathbf{x}\|}\right]_+\mathbf{x}$	Euclidean norm, $\lambda>0$
$\lambda\|\mathbf{x}\|_1$	$\mathbb{R}^n$	$[\lvert\mathbf{x}\rvert-\lambda\mathbf{e}]_+\circ\mathrm{sgn}(\mathbf{x})$	$\lambda>0$
$-\lambda\sum_{j=1}^n\log x_j$	$\mathbb{R}^n_{++}$	$\left(\frac{x_j+\sqrt{x_j^2+4\lambda}}{2}\right)_{j=1}^n$	$\lambda>0$
$\delta_C(\mathbf{x})$	$\mathbb{E}$	$P_C(\mathbf{x})$	$C\subseteq\mathbb{E}$
$\lambda\sigma_C(\mathbf{x})$	$\mathbb{E}$	$\mathbf{x}-\lambda P_C(\mathbf{x}/\lambda)$	C closed and convex
$\lambda\|\mathbf{x}\|$	$\mathbb{E}$	$\mathbf{x}-\lambda P_{B_{\|\cdot\|_*}[\mathbf{0},1]}(\mathbf{x}/\lambda)$	arbitrary norm
$\lambda\max\{x_1,x_2,\ldots,x_n\}$	$\mathbb{R}^n$	$\mathbf{x}-\mathrm{prox}_{\Delta_n}(\mathbf{x}/\boldsymbol{\lambda})$	$\lambda>0$
$\lambda d_C(\mathbf{x})$	$\mathbb{E}$	$\mathbf{x}+\min\left\{\frac{\lambda}{d_C(\mathbf{x})},1\right\}(P_C(\mathbf{x})-\mathbf{x})$	C closed convex
$\frac{\lambda}{2}d_C(\mathbf{x})^2$	$\mathbb{E}$	$\frac{\lambda}{\lambda+1}P_C(\mathbf{x})+\frac{1}{\lambda+1}\mathbf{x}$	C closed convex

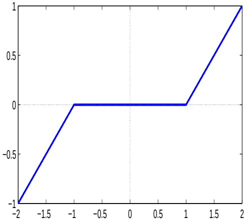
Prox of l_1-Norm

- $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$ $(\lambda > 0)$
- $g(\mathbf{x}) = \sum_{i=1}^n \varphi(x_i)$, where $\varphi(t) = \lambda|t|$.

- $\mathrm{prox}_{\varphi}(s) = \mathcal{T}_\lambda(s)$, where $\mathcal{T}_\lambda$ is defined as

$$\mathcal{T}_\lambda(y) = [|y| - \lambda]_+ \mathrm{sgn}(y) = \begin{cases} y - \lambda, & y \geq \lambda, \\ 0, & |y| < \lambda, \\ y + \lambda, & y \leq -\lambda \end{cases}$$

is the soft thresholding operator.



- By the separability of the l_1-norm, $\mathrm{prox}_g(\mathbf{x}) = (\mathcal{T}_\lambda(x_j))_{j=1}^n$. We expend the definition of the soft thresholding operator and write

$$\mathrm{prox}_g(\mathbf{x}) = \mathcal{T}_\lambda(\mathbf{x}) \equiv (\mathcal{T}_\lambda(x_j))_{j=1}^n = [|\mathbf{x}| - \lambda\mathbf{e}]_+ \odot \mathrm{sgn}(\mathbf{x}).$$

The Second Prox Theorem

Theorem Let $g : \mathbb{E} \to (-\infty, \infty]$ be a proper, closed and convex function. Then

(i) $\mathbf{u} = \mathrm{prox}_g(\mathbf{x})$.

(ii) $\mathbf{x} - \mathbf{u} \in \partial g(\mathbf{u})$.

(iii) $g(\mathbf{y}) \geq g(\mathbf{u}) + \langle \mathbf{x} - \mathbf{u}, \mathbf{y} - \mathbf{u} \rangle$ for any $\mathbf{y} \in \mathbb{E}$.

Proof.

- (i) is satisfied if and only if $\mathbf{u}$ a minimizer of the problem

$$\min_{\mathbf{u}} \left\{ g(\mathbf{u}) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2 \right\}.$$

- By Fermat's optimality condition, this is equivalent to (ii).
- The equivalence to (iii) follows by the definition of the subgradient.

Generalization of the second projection theorem!

Corollary: $\mathbf{x}$ is a minimizer of a closed, proper, convex function f iff $\mathbf{x} = \mathrm{prox}_f(\mathbf{x})$

Firm Nonexpansivity of the Prox Operator

Theorem. For any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$

(i) $\langle \mathbf{x} - \mathbf{y}, \mathrm{prox}_h(\mathbf{x}) - \mathrm{prox}_h(\mathbf{y}) \rangle \geq \|\mathrm{prox}_h(\mathbf{x}) - \mathrm{prox}_h(\mathbf{y})\|^2$.

(ii) $\|\mathrm{prox}_h(\mathbf{x}) - \mathrm{prox}_h(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|$.

Proof.

- Denote $\mathbf{u} = \mathrm{prox}_h(\mathbf{x}), \mathbf{v} = \mathrm{prox}_h(\mathbf{y})$.
- $\mathbf{x} - \mathbf{u} \in \partial h(\mathbf{u}), \mathbf{y} - \mathbf{v} \in \partial h(\mathbf{v})$.
- By the subgradient inequality

$$\begin{aligned} f(\mathbf{v}) &\geq f(\mathbf{u}) + \langle \mathbf{x} - \mathbf{u}, \mathbf{v} - \mathbf{u} \rangle, \\ f(\mathbf{u}) &\geq f(\mathbf{v}) + \langle \mathbf{y} - \mathbf{v}, \mathbf{u} - \mathbf{v} \rangle. \end{aligned}$$

- Summing the above two inequalities, we obtain $\langle (\mathbf{x} - \mathbf{u}) - (\mathbf{y} - \mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \geq 0$.
- Thus,$\langle \mathbf{u} - \mathbf{v}, \mathbf{x} - \mathbf{y} \rangle \geq \|\mathbf{u} - \mathbf{v}\|^2$.
- (ii) follows from Cauchy-Schwarz.

Moreau Decomposition

Theorem. Let f be a closed, proper and extended real-valued convex function. Then for any $\mathbf{x} \in \mathbb{E}$

$$\mathrm{prox}_f(\mathbf{x}) + \mathrm{prox}_{f^*}(\mathbf{x}) = \mathbf{x}.$$

Proof.

- Let $\mathbf{x} \in \mathbb{E}, \mathbf{u} = \mathrm{prox}_f(\mathbf{x})$.
- $\mathbf{x} - \mathbf{u} \in \partial f(\mathbf{u})$
- iff $\mathbf{u} \in \partial f^*(\mathbf{x} - \mathbf{u})$.
- iff $\mathbf{x} - \mathbf{u} = \mathrm{prox}_{f^*}(\mathbf{x})$.
- Thus,
$$\mathrm{prox}_f(\mathbf{x}) + \mathrm{prox}_{f^*}(\mathbf{x}) = \mathbf{u} + (\mathbf{x} - \mathbf{u}) = \mathbf{x}.$$

A direct consequence (extended Moreau decomposition)

$$\mathrm{prox}_{\lambda f}(\mathbf{x}) + \lambda \mathrm{prox}_{f^*/\lambda}(\mathbf{x}/\lambda) = \mathbf{x}$$

Prox of Support Functions

Let C be a nonempty closed and convex set, and let $\lambda > 0$. Then

$$\mathrm{prox}_{\lambda\sigma_C}(\mathbf{x}) = \mathbf{x} - \lambda P_C(\mathbf{x}/\lambda).$$

Proof. By the extended Moreau decomposition formula

$$\mathrm{prox}_{\lambda\sigma_C}(\mathbf{x}) = \mathbf{x} - \lambda\mathrm{prox}_{\lambda^{-1}\sigma_C^*}(\mathbf{x}/\lambda) = \mathbf{x} - \lambda\mathrm{prox}_{\lambda^{-1}\delta_C}(\mathbf{x}/\lambda) = \mathbf{x} - \lambda P_C(\mathbf{x}/\lambda)$$

Examples:

- $\mathrm{prox}_{\lambda\|\cdot\|_\alpha}(\mathbf{x}) = \mathbf{x} - \lambda P_{B_{\|\cdot\|_{\alpha,*}}[\mathbf{0},1]}(\mathbf{x}/\lambda)$. ($\|\cdot\|_\alpha$ - arbitrary norm)
- $\mathrm{prox}_{\lambda\|\cdot\|_\infty}(\mathbf{x}) = \mathbf{x} - \lambda P_{B_{\|\cdot\|_1}[\mathbf{0},1]}(\mathbf{x}/\lambda)$.
- $\mathrm{prox}_{\lambda\max(\cdot)}(\mathbf{x}) = \mathbf{x} - \lambda P_{\Delta_n}(\mathbf{x}/\lambda)$.

The Proximal Gradient Method

- A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci. (2009).
- A. Beck and M. Teboulle, *Gradient-based algorithms with applications to signal-recovery problems*, In Convex optimization in signal processing and communications (2010)
- H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces* (2011).
- P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward backward splitting*, Multiscale Model. Simul. (2005).
- N. Parikh and S. Boyd, *Proximal algorithms*, Foundations and Trends in Optimization (2014).
- J. Nutini, M. Schmidt, I. H. Laradji, M. Friendlander, and H. Koepke, *Coordinate descent converges faster with the gauss-southwell rule than random selection*, 32nd International Conference on Machine Learning (2015).

Preliminaries – Smoothness

Definition. Let $L \geq 0$. A function $f : \mathbb{E} \to (-\infty, \infty]$ is said to be L-smooth over a set $D \subseteq \text{int}(\text{dom}(f))$ if it is differentiable over D and satisfies

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x} - \mathbf{y}\| \text{ for all } \mathbf{x}, \mathbf{y} \in D.$$

The constant L is called the smoothness parameter.

- We consider here also non-Euclidean norms.
- The class of L-smooth functions is denoted by $C_L^{1,1}(D)$.
- When $D = \mathbb{E}$, the class is often denoted by $C_L^{1,1}$.
- The class of functions which are L-smooth for some $L \geq 0$ is denoted by $C^{1,1}$.
- If a function is L_1-smooth, then it is also L_2-smooth for any $L_2 \geq L_1$.

Examples:

- $f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + \mathbf{b}$, $\mathbf{a} \in \mathbb{E}$, $b \in \mathbb{R}$ (0-smooth).
- $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$, $\mathbf{A} \in \mathbb{S}^n, \mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$ ($\|\mathbf{A}\|_{p,q}$-smooth if $\mathbb{R}^n$ is endowed with the l_p-norm).
- $f(\mathbf{x}) = \frac{1}{2}d_C^2$ ($f : \mathbb{E} \to \mathbb{R}$) (1-smooth)

The Descent Lemma

Lemma. Let $f : \mathbb{E} \to (-\infty, \infty]$ be an L-smooth function ($L \geq 0$) over a given convex set D. Then for any $\mathbf{x}, \mathbf{y} \in D$,

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2.$$

Proof.

- By the fundamental theorem of calculus: $f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle dt$.
- $f(\mathbf{y}) - f(\mathbf{x}) = \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle dt$.
- Thus,

$$\begin{aligned}
|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| &= \left| \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle dt \right| \\
&\overset{(*)}{\leq} \int_0^1 \|\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\|_* \cdot \|\mathbf{y} - \mathbf{x}\| dt \\
&\leq \int_0^1 tL \|\mathbf{y} - \mathbf{x}\|^2 dt = \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2,
\end{aligned}$$

Characterizations of L-smoothness

Theorem. Let $f : \mathbb{E} \to \mathbb{R}$ be a convex function, differentiable over $\mathbb{E}$, and let $L > 0$. Then the following claims are equivalent:

(i) f is L-smooth.

(ii) $f(\mathbf{y}) \le f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{E}$.

(iii) $f(\mathbf{y}) \ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*^2$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{E}$.

(iv) $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \ge \frac{1}{L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*^2$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{E}$.

(v) $f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \ge \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) - \frac{L}{2}\lambda(1-\lambda)\|\mathbf{x} - \mathbf{y}\|^2$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$ and $\lambda \in [0, 1]$.

L-Smoothness and Boundedness of the Hessian

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function over $\mathbb{R}^n$. Then for a given $L \geq 0$, the following two claims are equivalent:

(i) f is L-smooth w.r.t. the l_p norm ($p \geq 1$).

(ii) $\|\nabla^2 f(\mathbf{x})\|_{p,q} \leq L$ for any $\mathbf{x} \in \mathbb{R}^n$, where q satisfies $\frac{1}{p} + \frac{1}{q} = 1$.

Corollary. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable convex function over $\mathbb{R}^n$. Then f is L-smooth w.r.t. the l_2-norm iff $\lambda_{\max}(\nabla^2 f(\mathbf{x})) \leq L$ for any $\mathbf{x} \in \mathbb{R}^n$.

Examples

- $f(\mathbf{x}) = \sqrt{1 + \|\mathbf{x}\|_2^2}$ ($f : \mathbb{R}^n \to \mathbb{R}$). 1-smooth w.r.t. to l_2.
- $f(\mathbf{x}) = \log\left(e^{x_1} + e^{x_2} + \ldots + e^{x_n}\right)$ ($f : \mathbb{R}^n \to \mathbb{R}$). 1-smooth w.r.t. l_2 and l_∞-norms.

The Proximal Gradient Method (PGM)

The Proximal Gradient Method aims to solve the composite model:

$$\text{(P)} \quad \min\{F(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}$$

(A) $g : \mathbb{E} \to (-\infty, \infty]$ is proper closed and convex.
(B) $f : \mathbb{E} \to (-\infty, \infty]$ is proper and closed; $\text{dom}(g) \subseteq \text{int}(\text{dom}(f))$ and f L_f-smooth over $\text{int}(\text{dom}(f))$.
(C) The optimal set of problem (P) is nonempty and denoted by X^*. The optimal value of the problem is denoted by F_{opt}.

Three prototype examples:

- **unconstrained smooth minimization** ($g \equiv 0$)
$$\min\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}$$
- **convex constrained smooth minimization** ($g = \delta_C, C \neq \emptyset$ closed convex)
$$\min\{f(\mathbf{x}) : \mathbf{x} \in C\}$$
- l_1 **regularized problems** ($\mathbb{E} = \mathbb{R}^n$, $g(x) \equiv \lambda\|x\|_1$)
$$\min\{f(\mathbf{x}) + \lambda\|\mathbf{x}\|_1 : \mathbf{x} \in \mathbb{R}^n\}$$

The Idea

Instead of minimizing directly

$$\min_{\mathbf{x}\in\mathbb{E}} f(\mathbf{x}) + g(\mathbf{x})$$

Approximate f by a regularized linear approximation of f while keeping g fixed.

$$\mathbf{x}^{k+1} = \operatorname*{argmin}_{\mathbf{x}} \left\{ f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k) + \frac{1}{2t_k}\|\mathbf{x} - \mathbf{x}^k\|^2 + g(\mathbf{x}) \right\}$$

$$\mathbf{x}^{k+1} = \operatorname*{argmin}_{\mathbf{x}} \left\{ g(\mathbf{x}) + \frac{1}{2t_k}\left\|\mathbf{x} - (\mathbf{x}^k - t_k\nabla f(\mathbf{x}^k))\right\|^2 \right\}$$

Proximal Gradient Method

$$\mathbf{x}^{k+1} = \mathrm{prox}_{t_k g}(\mathbf{x}^k - t_k\nabla f(\mathbf{x}^k))$$

Three Prototype Examples Contd.

- **Gradient Method** ($g = 0$, unconstrained minimization)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - t_k \nabla f(\mathbf{x}^k)$$

- **Gradient Projection Method** ($g = \delta_C$, constrained convex minimization)

$$\mathbf{x}^{k+1} = P_C(\mathbf{x}^k - t_k \nabla f(\mathbf{x}^k))$$

- **Iterative Soft-Thresholding Algorithm (ISTA)** ($g = \|\cdot\|_1$):

$$\mathbf{x}^{k+1} = \mathcal{T}_{\lambda t_k}\left(\mathbf{x}^k - t_k \nabla f(\mathbf{x}^k)\right)$$

where $\mathcal{T}_\alpha(\mathbf{u}) = [|\mathbf{u}| - \alpha \mathbf{e}] \odot \mathrm{sgn}(\mathbf{u})$.

The Proximal Gradient Method

- We will take the stepsizes as $t_k = \frac{1}{L_k}$.

The Proximal Gradient Method

Initialization: pick $\mathbf{x}^0 \in \mathrm{int}(\mathrm{dom}(f))$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick $L_k > 0$.

(b) set $\mathbf{x}^{k+1} = \mathrm{prox}_{\frac{1}{L_k}g}\left(\mathbf{x}^k - \frac{1}{L_k}\nabla f(\mathbf{x}^k)\right)$.

- The general update step can be written as $\mathbf{x}^{k+1} = T_{L_k}^{f,g}(\mathbf{x}^k)$
- $T_L^{f,g} : \mathrm{int}(\mathrm{dom}(f)) \to \mathbb{E}$ is the prox-grad operator defined by

$$T_L^{f,g}(\mathbf{x}) \equiv \mathrm{prox}_{\frac{1}{L}g}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right).$$

- When the identities of f and g will be clear from the context, we will often omit the superscripts f, g and write $T_L(\cdot)$ instead of $T_L^{f,g}(\cdot)$.

Sufficient Decrease Lemma

Lemma. Let $F = f + g$ and $T_L \equiv T_L^{f,g}$. Then for any $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$ and $L \in \left(\frac{L_f}{2}, \infty\right)$

$$F(\mathbf{x}) - F(T_L(\mathbf{x})) \geq \frac{L - \frac{L_f}{2}}{L^2} \left\| G_L^{f,g}(\mathbf{x}) \right\|^2, \tag{4}$$

where $G_L^{f,g} : \mathrm{int}(\mathrm{dom}(f)) \to \mathbb{E}$ is the operator defined by $G_L^{f,g}(\mathbf{x}) = L(\mathbf{x} - T_L(\mathbf{x}))$.

Proof. We use the shorthand notation $\mathbf{x}^+ = T_L(\mathbf{x})$.

- By the descent lemma

$$f(\mathbf{x}^+) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}^+ - \mathbf{x} \rangle + \frac{L_f}{2} \|\mathbf{x} - \mathbf{x}^+\|^2. \tag{5}$$

- By the second prox theorem, since $\mathbf{x}^+ = \mathrm{prox}_{\frac{1}{L}g}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right)$,

$$\left\langle \mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x}) - \mathbf{x}^+, \mathbf{x} - \mathbf{x}^+ \right\rangle \leq \frac{1}{L}g(\mathbf{x}) - \frac{1}{L}g(\mathbf{x}^+).$$

- Thus, $\langle \nabla f(\mathbf{x}), \mathbf{x}^+ - \mathbf{x} \rangle \leq -L \left\|\mathbf{x}^+ - \mathbf{x}\right\|^2 + g(\mathbf{x}) - g(\mathbf{x}^+)$,
- which combined with (5) yields

$$f(\mathbf{x}^+) + g(\mathbf{x}^+) \leq f(\mathbf{x}) + g(\mathbf{x}) + \left(-L + \frac{L_f}{2}\right) \left\|\mathbf{x}^+ - \mathbf{x}\right\|^2.$$

The Gradient Mapping

- Definition. The gradient mapping is the operator $G_L^{f,g} : \mathrm{int}(\mathrm{dom}(f)) \to \mathbb{E}$ defined by
$$G_L^{f,g}(\mathbf{x}) \equiv L\left(\mathbf{x} - T_L^{f,g}(\mathbf{x})\right)$$
for any $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$.
- When the identities of f and g will be clear from the context, we will use the notation G_L instead of $G_L^{f,g}$.

In the special case where $L = L_f$, the sufficient decrease lemma amounts to

Corollary. For any $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$:
$$F(\mathbf{x}) - F(T_{L_f}(\mathbf{x})) \geq \frac{1}{2L_f} \left\| G_{L_f}(\mathbf{x}) \right\|^2.$$

Properties of the Gradient Mapping I

Recall: under properties (A),(B), the stationary points of the problem

$$(P) \quad \min\{F(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x})\}$$

are the points satisfying $-\nabla f(\mathbf{x}) \in \partial g(\mathbf{x})$. Necessary optimality condition when f is nonconvex, and necessary and sufficient condition if f is convex.

Theorem Let f and g satisfy properties (A) and (B) and let $L > 0$. Then

(a) $G_L^{f,g_0}(\mathbf{x}) = \nabla f(\mathbf{x})$ for any $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$, where $g_0(\mathbf{x}) \equiv 0$.

(b) For $\mathbf{x}^* \in \mathrm{int}(\mathrm{dom}(f))$, $G_L^{f,g}(\mathbf{x}^*) = \mathbf{0}$ iff $\mathbf{x}^*$ is a stationary point

Proof.

(a) $G_L^{f,g_0}(\mathbf{x}) = L\left(\mathbf{x} - \mathrm{prox}_{\frac{1}{L}g_0}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right)\right) = L\left(\mathbf{x} - \left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right)\right) = \nabla f(\mathbf{x})$.

(b) $G_L^{f,g}(\mathbf{x}^*) = \mathbf{0}$ iff $\mathbf{x}^* = \mathrm{prox}_{\frac{1}{L}g}\left(\mathbf{x}^* - \frac{1}{L}\nabla f(\mathbf{x}^*)\right)$. By the second prox theorem

$$\mathbf{x}^* - \frac{1}{L}\nabla f(\mathbf{x}^*) - \mathbf{x}^* \in \frac{1}{L}\partial g(\mathbf{x}^*),$$

that is, iff $-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*)$.

The Gradient Mapping as an Optimality Measure

Corollary Let f and g satisfy properties (A) and (B) and let $L > 0$. Suppose that in addition f is convex. Then for $\mathbf{x}^* \in \text{dom}(g)$, $G_L^{f,g}(\mathbf{x}^*) = \mathbf{0}$ if and only if $\mathbf{x}^*$ is an optimal solution of problem (P).

- $\|G_L(\mathbf{x})\|$ can be regarded as an "optimality measure" in the sense that it is always nonnegative, and equal to zero if and only if $\mathbf{x}$ is a stationary point (or optimal point if f is convex).

Properties of the Gradient Mapping II

- **monotonicity w.r.t. the parameter.** for any $\mathbf{x} \in \text{int}(\text{dom}(f))$ and $L_1 \geq L_2 > 0$,

$$\begin{aligned} \|G_{L_1}(\mathbf{x})\| &\geq \|G_{L_2}(\mathbf{x})\|, \\ \frac{\|G_{L_1}(\mathbf{x})\|}{L_1} &\leq \frac{\|G_{L_2}(\mathbf{x})\|}{L_2}. \end{aligned}$$

- **Lipschitz continuity.** $\|G_L(\mathbf{x}) - G_L(\mathbf{y})\| \leq (2L + L_f)\|\mathbf{x} - \mathbf{y}\|$.

If in addition f is convex and L_f-smooth (over the entire space)

- $\langle G_{L_f}(\mathbf{x}) - G_{L_f}(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \geq \frac{3}{4L_f}\|G_{L_f}(\mathbf{x}) - G_{L_f}(\mathbf{y})\|^2$
- $\|G_{L_f}(\mathbf{x}) - G_{L_f}(\mathbf{y})\| \leq \frac{4L_f}{3}\|\mathbf{x} - \mathbf{y}\|$
- **Monotonicity w.r.t. the prox-grad mapping:** $\|G_{L_f}(T_{L_f}(\mathbf{x}))\| \leq \|G_{L_f}(\mathbf{x})\|$.

Stepsize Strategies

- **constant.** $L_k = \bar{L} \in \left(\frac{L_f}{2}, \infty\right)$ for all k.
- **backtracking procedure B1.** The procedure requires three parameters (s, γ, η) where $s > 0, \gamma \in (0,1)$ and $\eta > 1$. First, L_k is set to be equal to the initial guess s. Then, while

$$F(\mathbf{x}^k) - F(T_{L_k}(\mathbf{x}^k)) < \frac{\gamma}{L_k}\|G_{L_k}(\mathbf{x}^k)\|^2,$$

we set $L_k := \eta L_k$. That is, L_k is chosen as $L_k = s\eta^{i_k}$, where i_k is the smallest nonnegative integer for which the condition

$$F(\mathbf{x}^k) - F(T_{s\eta^{i_k}}(\mathbf{x}^k)) \geq \frac{\gamma}{s\eta^{i_k}}\|G_{s\eta^{i_k}}(\mathbf{x}^k)\|^2$$

is satisfied.

For the backtracking procedure it holds that $L_k \leq \max\left\{s, \frac{\eta L_f}{2(1-\gamma)}\right\}$.

Sufficient Decrease For Proximal Gradient

Lemma. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by PGM. with either a constant stepsize defined by $L_k = \bar{L} \in \left(\frac{L_f}{2}, \infty\right)$ or with a stepsize chosen by the backtracking procedure B1. Then

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq M\|G_d(\mathbf{x}^k)\|^2,$$

where

$$M = \begin{cases} \frac{\bar{L} - \frac{L_f}{2}}{(\bar{L})^2} & \text{constant stepsize,} \\ \frac{\gamma}{\max\left\{s, \frac{\eta L_f}{2(1-\gamma)}\right\}} & \text{backtracking,} \end{cases} \qquad d = \begin{cases} \bar{L}, & \text{constant stepsize,} \\ s, & \text{backtracking.} \end{cases}$$

Proof. The result for the constant stepsize setting follows by plugging $L = \bar{L}$ and $\mathbf{x} = \mathbf{x}^k$ in the sufficient decrease lemma. For the backtracking procedure we have

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq \frac{\gamma}{L_k}\|G_{L_k}(\mathbf{x}^k)\|^2 \geq \frac{\gamma}{\max\left\{s, \frac{\eta L_f}{2(1-\gamma)}\right\}}\|G_{L_k}(\mathbf{x}^k)\|^2 \geq \frac{\gamma}{\max\left\{s, \frac{\eta L_f}{2(1-\gamma)}\right\}}\|G_s(\mathbf{x}^k)\|^2,$$

Convergence of PGM - the Nonconvex Case

Theorem. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by PGM with either a constant stepsize defined by $L_k = \bar{L} \in \left(\frac{L_f}{2}, \infty\right)$ or with a stepsize chosen by the backtracking procedure B1. Then

(a) The sequence $\{F(\mathbf{x}^k)\}_{k\geq 0}$ is nonincreasing. In addition, $F(\mathbf{x}^{k+1}) < F(\mathbf{x}^k)$ if and only if $\mathbf{x}^k$ is not a stationary point of (P).

(b) $G_d(\mathbf{x}^k) \to \mathbf{0}$ as $k \to \infty$.

(c) $\min_{n=0,1,\ldots,k} \|G_d(\mathbf{x}^n)\| \leq \frac{\sqrt{F(\mathbf{x}^0) - F_{\rm opt}}}{\sqrt{M(k+1)}}$.

(d) All limit points of the sequence $\{\mathbf{x}^k\}_{k\geq 0}$ are stationary points of problem (P).

The Fundamental Prox-Grad Inequality

Theorem. For any $\mathbf{x} \in \mathbb{E}$ and $\mathbf{y} \in \mathrm{int}(\mathrm{dom}(f))$ satisfying

$$f(T_L(\mathbf{y})) \le f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), T_L(\mathbf{y}) - \mathbf{y} \rangle + \frac{L}{2}\|T_L(\mathbf{y}) - \mathbf{y}\|^2, \tag{6}$$

it holds that

$$F(\mathbf{x}) - F(T_L(\mathbf{y})) \ge \frac{L}{2}\|\mathbf{x} - T_L(\mathbf{y})\|^2 - \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 + \ell_f(\mathbf{x}, \mathbf{y}), \tag{7}$$

where $\ell_f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$.

Proof.

- We use the notation $\mathbf{y}^+ = T_L(\mathbf{y})$.
- Since $\mathbf{y}^+ = \mathrm{prox}_{\frac{1}{L}g}\left(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\right)$, by the second prox theorem it follows that

$$\frac{1}{L}g(\mathbf{x}) \ge \frac{1}{L}g(\mathbf{y}^+) + \left\langle \mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y}) - \mathbf{y}^+, \mathbf{x} - \mathbf{y}^+ \right\rangle.$$

- Therefore,

$$\begin{aligned} g(\mathbf{x}) &\ge g(\mathbf{y}^+) + L\langle \mathbf{y} - \mathbf{y}^+, \mathbf{x} - \mathbf{y}^+ \rangle + \langle \nabla f(\mathbf{y}), \mathbf{y}^+ - \mathbf{x} \rangle \\ &= g(\mathbf{y}^+) + L\langle \mathbf{y} - \mathbf{y}^+, \mathbf{x} - \mathbf{y}^+ \rangle \\ &\quad + \langle \nabla f(\mathbf{y}), \mathbf{y}^+ - \mathbf{y} \rangle + \langle \nabla f(\mathbf{y}), \mathbf{y} - \mathbf{x} \rangle \qquad (8) \end{aligned}$$

Proof Contd.

- By (6), $f(\mathbf{y}^+) \le f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{y}^+ - \mathbf{y} \rangle + \frac{L}{2}\|\mathbf{y}^+ - \mathbf{y}\|^2$
- Hence, $\langle \nabla f(\mathbf{y}), \mathbf{y}^+ - \mathbf{y} \rangle \ge f(\mathbf{y}^+) - f(\mathbf{y}) - \frac{L}{2}\|\mathbf{y}^+ - \mathbf{y}\|^2$,
- which combined with (8) yields

$$F(\mathbf{x}) \ge F(\mathbf{y}^+) + L\langle \mathbf{y} - \mathbf{y}^+, \mathbf{x} - \mathbf{y}^+ \rangle - \frac{L}{2}\|\mathbf{y}^+ - \mathbf{y}\|^2 + \ell_f(\mathbf{x}, \mathbf{y}).$$

- Using the identity $\langle \mathbf{y} - \mathbf{y}^+, \mathbf{x} - \mathbf{y}^+ \rangle = \frac{1}{2}\|\mathbf{x} - \mathbf{y}^+\|^2 + \frac{1}{2}\|\mathbf{y} - \mathbf{y}^+\|^2 - \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2$, we obtain that

$$F(\mathbf{x}) - F(\mathbf{y}^+) \ge \frac{L}{2}\|\mathbf{x} - \mathbf{y}^+\|^2 - \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 + \ell_f(\mathbf{x}, \mathbf{y}),$$

Sufficient Decrease Lemma - 2nd Version

Corollary. For any $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$ for which

$$f(T_L(\mathbf{x})) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), T_L(\mathbf{x}) - \mathbf{x} \rangle + \frac{L}{2} \| T_L(\mathbf{x}) - \mathbf{x} \|^2,$$

it holds that

$$F(\mathbf{x}) - F(T_L(\mathbf{x})) \geq \frac{1}{2L} \| G_L(\mathbf{x}) \|^2.$$

Stepsize Strategies in the Convex Case

When f is also convex, we will define two possible stepsize strategies for which

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L_k}{2} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2.$$

- **constant.** $L_k = L_f$ for all k.
- **backtracking procedure B2.** The procedure requires two parameters (s, η), where $s > 0$ and $\eta > 1$. Define $L_{-1} = s$. At iteration k, L_k is set to be equal to L_{k-1}. Then, while

$$f(T_{L_k}(\mathbf{x}^k)) > f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), T_{L_k}(\mathbf{x}^k) - \mathbf{x}^k \rangle + \frac{L_k}{2} \| T_{L_k}(\mathbf{x}^k) - \mathbf{x}^k\|^2,$$

we set $L_k := \eta L_k$. That is, L_k is chosen as $L_k = L_{k-1}\eta^{i_k}$, where i_k is the smallest nonnegative integer for which

$$f(T_{L_{k-1}\eta^{i_k}}(\mathbf{x}^k)) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), T_{L_{k-1}\eta^{i_k}}(\mathbf{x}^k) - \mathbf{x}^k \rangle + \frac{L_k}{2} \| T_{L_{k-1}\eta^{i_k}}(\mathbf{x}^k) - \mathbf{x}^k\|^2.$$

Remarks

- $\beta L_f \le L_k \le \alpha L_f$, where

$$\alpha = \begin{cases} 1, & \text{constant}, \\ \max\left\{\eta, \frac{s}{L_f}\right\}, & \text{backtracking}, \end{cases} \quad \beta = \begin{cases} 1, & \text{constant}, \\ \frac{s}{L_f}, & \text{backtracking}. \end{cases}$$

- **Monotonicity of PGM.** Invoking the sufficient decrease lemma (2nd version) with $\mathbf{x} = \mathbf{x}^k$, we obtain that

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \ge \frac{L_k}{2}\|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2.$$

or

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \ge \frac{1}{2L_k}\|G_{L_k}(\mathbf{x}^k)\|^2.$$

$O(1/k)$ Rate of Convergence of Proximal Gradient

Theorem. Suppose that f is convex. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method with either a constant stepsize rule or the backtracking procedure B2. Then for any $\mathbf{x}^* \in X^*$ and $k \geq 0$,

$$F(\mathbf{x}^k) - F_{\rm opt} \leq \frac{\alpha L_f \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2k},$$

where $\alpha = 1$ in the constant stepsize setting and $\alpha = \max\left\{\eta, \frac{s}{L_f}\right\}$ if the backtracking rule is employed.

Proof.

- Substituting $L = L_n$, $\mathbf{x} = \mathbf{x}^*$ and $\mathbf{y} = \mathbf{x}^n$ in the fundamental prox-grad ineq.,

$$\begin{aligned}\frac{2}{L_n}(F(\mathbf{x}^*) - F(\mathbf{x}^{n+1})) &\geq \|\mathbf{x}^* - \mathbf{x}^{n+1}\|^2 - \|\mathbf{x}^* - \mathbf{x}^n\|^2 + \frac{2}{L_n}\ell_f(\mathbf{x}^*, \mathbf{x}^n)\\ &\geq \|\mathbf{x}^* - \mathbf{x}^{n+1}\|^2 - \|\mathbf{x}^* - \mathbf{x}^n\|^2,\end{aligned}$$

Proof Contd.

- Summing over $n = 0, 1, \ldots, k-1$ and using the bound $L_n \le \alpha L_f$, we obtain

$$\frac{2}{\alpha L_f} \sum_{n=0}^{k-1} (F(\mathbf{x}^*) - F(\mathbf{x}^{n+1})) \ge \|\mathbf{x}^* - \mathbf{x}^k\|^2 - \|\mathbf{x}^* - \mathbf{x}^0\|^2.$$

- $\sum_{n=0}^{k-1}(F(\mathbf{x}^{n+1}) - F_{\text{opt}}) \le \frac{\alpha L_f}{2}\|\mathbf{x}^* - \mathbf{x}^0\|^2 - \frac{\alpha L_f}{2}\|\mathbf{x}^* - \mathbf{x}^k\|^2 \le \frac{\alpha L_f}{2}\|\mathbf{x}^* - \mathbf{x}^0\|^2.$
- By the monotonicity of $\{F(\mathbf{x}^n)\}_{n\ge 0}$,

$$k(F(\mathbf{x}^k) - F_{\text{opt}}) \le \sum_{n=0}^{k-1}(F(\mathbf{x}^{n+1}) - F_{\text{opt}}) \le \frac{\alpha L_f}{2}\|\mathbf{x}^* - \mathbf{x}^0\|^2.$$

- Consequently, $F(\mathbf{x}^k) - F_{\text{opt}} \le \frac{\alpha L_f \|\mathbf{x}^* - \mathbf{x}^0\|^2}{2k}$.

Fejér Monotonicity

Theorem. Suppose that f is convex. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method with either a constant stepsize rule or the backtracking procedure B2. Then for any $\mathbf{x}^* \in X^*$ and $k \geq 0$,

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

Proof.

- Substituting $L = L_k$, $\mathbf{x} = \mathbf{x}^*$ and $\mathbf{y} = \mathbf{x}^k$ in the fundamental prox-grad inequality (7),

$$\begin{aligned} \frac{2}{L_k}(F(\mathbf{x}^*) - F(\mathbf{x}^{k+1})) &\geq \|\mathbf{x}^* - \mathbf{x}^{k+1}\|^2 - \|\mathbf{x}^* - \mathbf{x}^k\|^2 + \frac{2}{L_k}\ell_f(\mathbf{x}^*, \mathbf{x}^k) \\ &\geq \|\mathbf{x}^* - \mathbf{x}^{k+1}\|^2 - \|\mathbf{x}^* - \mathbf{x}^k\|^2, \end{aligned}$$

- The result follows by the inequality $F(\mathbf{x}^*) - F(\mathbf{x}^{k+1}) \leq 0$.

Fejér Monotonicity - Definition and Main Result

- Definition. A sequence $\{\mathbf{x}^k\}_{k\geq 0} \subseteq \mathbb{E}$ is called Fejér monotone w.r.t. a set $S \subseteq \mathbb{E}$ if $\|\mathbf{x}^{k+1} - \mathbf{y}\| \leq \|\mathbf{x}^k - \mathbf{y}\|$ for all $k \geq 0$ and $\mathbf{y} \in S$.

Theorem (convergence of Fejér monotone sequences). Let $\{\mathbf{x}^k\}_{k\geq 0} \subseteq \mathbb{E}$ be asequence, and let S be a set satisfying $D \subseteq S$, where D is the set comprising all the limit points of $\{\mathbf{x}^k\}_{k\geq 0}$. If $\{\mathbf{x}^k\}_{k\geq 0}$ is Fejér monotone w.r.t. S, then it converges to a point in D.

Consequence: convergence of the sequence generated by PGM.

Theorem. Suppose that f is convex. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by PGM with either a constant stepsize rule or the backtracking procedure B2. Then the sequence $\{\mathbf{x}^k\}_{k\geq 0}$ converges to an optimal solution of problem (P).

Iteration Complexity of Algorithms

- An ε-optimal solution of problem (P) is a vector $\bar{\mathbf{x}} \in \text{dom}(g)$ satisfying $F(\bar{\mathbf{x}}) - F_{\text{opt}} \leq \varepsilon$.
- In complexity analysis, the following question is asked: how many iterations are required to obtain an ε-optimal solution? meaning how many iterations are required to obtain the condition $F(\mathbf{x}^k) - F_{\text{opt}} \leq \varepsilon$
- Recall: $F(\mathbf{x}^k) - F_{\text{opt}} \leq \frac{\alpha L_f \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2k}$.

Theorem[$O(1/\varepsilon)$ complexity of PGM]. For any k satisfying

$$k \geq \left\lceil \frac{\alpha L_f R^2}{2\varepsilon} \right\rceil$$

it holds that $F(\mathbf{x}^k) - F_{\text{opt}} \leq \varepsilon$, where R is an upper bound on $\|\mathbf{x}^* - \mathbf{x}^0\|$ for some $\mathbf{x}^* \in X^*$.

$O(1/k)$ Rate of Convergence of the Gradient Mapping Norm in the Convex Case

Recall: $\min_{n=0,1,\ldots,k} \|G_d(\mathbf{x}^n)\| \leq \frac{\sqrt{F(\mathbf{x}^0)-F_{\rm opt}}}{\sqrt{M(k+1)}}$.

We can do better if f is convex:

Theorem. Suppose that f is convex. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by PGM with either a constant stepsize by the backtracking procedure B2. Then for any $\mathbf{x}^* \in X^*$ and $k \geq 0$,

$$\min_{n=0,1,\ldots,k} \|G_{\alpha L_f}(\mathbf{x}^n)\| \leq \frac{2\alpha^{1.5} L_f \|\mathbf{x}^0 - \mathbf{x}^*\|}{\sqrt{\beta}(k+1)}.$$

where $\alpha = \beta = 1$ in the constant stepsize setting and $\alpha = \max\left\{\eta, \frac{s}{L_f}\right\}, \beta = \frac{s}{L_f}$ if the backtracking rule is employed.

And even better if a constant stepsize is used: $\|G_{L_f}(\mathbf{x}^k)\| \leq \frac{2L_f\|\mathbf{x}^0-\mathbf{x}^*\|}{k+1}$.

Linear Rate of Convergence of PGM – Strongly Convex Case

Theorem. Suppose that f is σ-strongly convex ($\sigma > 0$). Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method with either a constant stepsize rule or backtracking procedure B2. Let

$$\alpha = \begin{cases} 1, & \text{constant stepsize}, \\ \max\left\{\eta, \frac{s}{L_f}\right\}, & \text{backtracking}. \end{cases}$$

Then for any $\mathbf{x}^* \in X$ and $k \geq 0$,

(a) $\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\sigma}{\alpha L_f}\right) \|\mathbf{x}^k - \mathbf{x}^*\|^2$.

(b) $\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\sigma}{\alpha L_f}\right)^k \|\mathbf{x}^0 - \mathbf{x}^*\|^2$.

(c) $F(\mathbf{x}^{k+1}) - F_{\rm opt} \leq \frac{\alpha L_f}{2} \left(1 - \frac{\sigma}{\alpha L_f}\right)^{k+1} \|\mathbf{x}^0 - \mathbf{x}^*\|^2$.

Complexity of PGM - the Strongly Convex Case

A direct result of the rate analysis:

Theorem. For any $k \geq 1$ satisfying

$$k \geq \alpha\kappa \log\left(\frac{1}{\varepsilon}\right) + \alpha\kappa \log\left(\frac{\alpha L_f R^2}{2}\right),$$

it holds that $F(\mathbf{x}^k) - F_{\rm opt} \leq \varepsilon$, where R is an upper bound on $\|\mathbf{x}^0 - \mathbf{x}^*\|$ and $\kappa = \frac{L_f}{\sigma}$.

Non-Euclidean Spaces

- Until now we assumed that the underlying space is Euclidean, meaning that $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$.
- What is the effect of considering a different norm?
- What is the role of the dual space?
- We will concentrate the simplest example: the gradient method.

The Dual Space

- A linear functional on a vector space $\mathbb{E}$ is a linear transformation from $\mathbb{E}$ to $\mathbb{R}$.
- The dual space $\mathbb{E}^*$ is the set of all linear functionals on $\mathbb{E}$.
- **Fact:** For inner product spaces, for any linear functional $f \in \mathbb{E}^*$, there exists $\mathbf{v} \in \mathbb{E}$ such that

$$f(\mathbf{x}) = \langle \mathbf{v}, \mathbf{x} \rangle.$$

- We will make the association $f(\cdot) = \langle \mathbf{v}, \cdot \rangle \in \mathbb{E}^* \leftrightarrow \mathbf{v} \in \mathbb{E}$.
- Convention: the elements in $\mathbb{E}^*$ are the same as in $\mathbb{E}$.
- The inner product in $\mathbb{E}^*$ is the same as in $\mathbb{E}$.
- Essentially, the only difference is the norm of the dual space:

$$\|\mathbf{y}\|_* \equiv \max_{\mathbf{x}}\{\langle \mathbf{y}, \mathbf{x} \rangle : \|\mathbf{x}\| \le 1\}, \quad \mathbf{y} \in \mathbb{E}^*.$$

- Alternative representation:

$$\|\mathbf{y}\|_* = \max_{\mathbf{x}}\{\langle \mathbf{y}, \mathbf{x} \rangle : \|\mathbf{x}\| = 1\}, \quad \mathbf{y} \in \mathbb{E}^*.$$

- Subgradients and gradients are always in the dual space.

Gradient Method Revisited

- Consider the unconstrained problem

$$\min\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\},$$

where we assume that f is L_f-smooth w.r.t. the underlying norm:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L_f \|\mathbf{x} - \mathbf{y}\|.$$

- The gradient method has the form

$$\mathbf{x}^{k+1} = \mathbf{x}^k - t_k \nabla f(\mathbf{x}^k).$$

- **A "philosophical" flaw:** $\mathbf{x}^k \in \mathbb{E}$ while $\nabla f(\mathbf{x}^k) \in \mathbb{E}^*$.
- **Solution:** consider the "primal counterpart" of $\nabla f(\mathbf{x}^k) \in \mathbb{E}^*$.

The Primal Counterpart

- Definition. For any vector $\mathbf{a} \in \mathbb{E}^*$, the set of primal counterparts of $\mathbf{a}$ is

$$\Lambda_{\mathbf{a}} = \underset{\mathbf{v} \in \mathbb{E}}{\operatorname{argmax}} \{\langle \mathbf{a}, \mathbf{v} \rangle : \|\mathbf{v}\| \leq 1\}.$$

Lemma [basic properties of primal counterparts] Let $\mathbf{a} \in \mathbb{E}^*$. Then

(a) If $\mathbf{a} \neq \mathbf{0}$, then $\|\mathbf{a}^\dagger\| = 1$ for any $\mathbf{a}^\dagger \in \Lambda_{\mathbf{a}}$.

(b) If $\mathbf{a} = \mathbf{0}$, then $\Lambda_{\mathbf{a}} = B_{\|\cdot\|}[\mathbf{0}, 1]$.

(c) $\langle \mathbf{a}, \mathbf{a}^\dagger \rangle = \|\mathbf{a}\|_*$ for any $\mathbf{a}^\dagger \in \Lambda_{\mathbf{a}}$.

Examples: $\mathbb{E} = \mathbb{R}^n$, $\mathbf{a} \neq \mathbf{0}$,

- $\|\cdot\| = \|\cdot\|_2$ $\Lambda_{\mathbf{a}} = \left\{\frac{\mathbf{a}}{\|\mathbf{a}\|_2}\right\}$.
- $\|\cdot\| = \|\cdot\|_1$, $\Lambda_{\mathbf{a}} = \left\{\sum_{i \in I(\mathbf{a})} \lambda_i \operatorname{sgn}(a_i)\mathbf{e}_i : \sum_{i \in I(\mathbf{a})} \lambda_i = 1, \lambda_j \geq 0, j \in I(\mathbf{a})\right\}$, where $I(\mathbf{a}) = \underset{i=1,2,\ldots,n}{\operatorname{argmax}} |a_i|$.
- $\|\cdot\| = \|\cdot\|_\infty$. $\Lambda_{\mathbf{a}} = \{\mathbf{z} \in \mathbb{R}^n : z_i = \operatorname{sgn}(a_i), i \in I_{\neq}(\mathbf{a}), |z_j| \leq 1, j \in I_0(\mathbf{a})\}$, where
$I_{\neq}(\mathbf{a}) = \{i \in \{1, 2, \ldots, n\} : a_i \neq 0\}, I_0(\mathbf{a}) = \{i \in \{1, 2, \ldots, n\} : a_i = 0\}$.

The Non-Euclidean Gradient Method

The Non-Euclidean Gradient Method

Initialization: pick $\mathbf{x}^0 \in \mathbb{E}$ arbitrarily.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick $\nabla f(\mathbf{x}^k)^\dagger \in \Lambda_{\nabla f(\mathbf{x}^k)}$;

(b) set $\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{\|\nabla f(\mathbf{x}^k)\|_*}{L_f} \nabla f(\mathbf{x}^k)^\dagger$.

- Convergence analysis relies on the descent lemma:
  $f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L_f}{2} \|\mathbf{x} - \mathbf{y}\|^2$.
- **Sufficient Decrease:** $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq \frac{1}{2L_f} \|\nabla f(\mathbf{x}^k)\|_*^2$.
- **Proof of sufficient decrease:**

$$
\begin{aligned}
f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L_f}{2} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \\
&= f(\mathbf{x}^k) - \frac{\|\nabla f(\mathbf{x}^k)\|_*}{L_f} \langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^k)^\dagger \rangle + \frac{\|\nabla f(\mathbf{x}^k)\|_*^2}{2L_f^2} \\
&= f(\mathbf{x}^k) - \frac{1}{2L_f} \|\nabla f(\mathbf{x}^k)\|_*^2,
\end{aligned}
$$

Convergence in the Nonconvex Case

Theorem. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the non-Euclidean gradient method. Then

(a) the sequence $\{f(\mathbf{x}^k)\}_{k\geq 0}$ is nonincreasing. In addition, $f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k)$ iff $\nabla f(\mathbf{x}^k) \neq \mathbf{0}$;

(b) if the sequence $\{f(\mathbf{x}^k)\}_{k\geq 0}$ is bounded below, then $\nabla f(\mathbf{x}^k) \to \mathbf{0}$ as $k \to \infty$;

(c) if the optimal value is finite and equal to $f_{\rm opt}$, then $\min_{n=0,1,\ldots,k} \|\nabla f(\mathbf{x}^n)\|_* \leq \frac{\sqrt{2L_f}\sqrt{f(\mathbf{x}^0)-f_{\rm opt}}}{\sqrt{k+1}}$.

(d) all limit points of the sequence $\{\mathbf{x}^k\}_{k\geq 0}$ are stationary points of f.

Proof. (a),(b) and (d) follow immediately from the sufficient decrease property. (c) follows by summing the sufficient decrease property

$$
\begin{aligned}
f(\mathbf{x}^0) - f_{\rm opt} &\geq f(\mathbf{x}^0) - f(\mathbf{x}^{k+1}) = \sum_{n=0}^{k}(f(\mathbf{x}^n) - f(\mathbf{x}^{n+1})) \\
&\geq \frac{1}{2L_f}\sum_{n=0}^{k}\|\nabla f(\mathbf{x}^n)\|_*^2 \geq \frac{k+1}{2L_f}\min_n \|\nabla f(\mathbf{x}^n)\|_*^2
\end{aligned}
$$

Convergence in the Convex Case

Assumptions:

- $f : \mathbb{E} \to \mathbb{R}$ is L_f-smooth and convex.
- The optimal set is nonempty and denoted by X^*. The optimal value is denoted by $f_{\rm opt}$.
- There exists $R > 0$ s.t. $\max_{\mathbf{x},\mathbf{x}^*}\{\|\mathbf{x}^* - \mathbf{x}\| : f(\mathbf{x}) \le f(\mathbf{x}^0), \mathbf{x}^* \in X^*\} \le R$.

Lemma. $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \ge \frac{1}{2L_f R^2}(f(\mathbf{x}^k) - f_{\rm opt})^2$

Proof.

- By the gradient inequality,

$$f(\mathbf{x}^k) - f_{\rm opt} = f(\mathbf{x}^k) - f(\mathbf{x}^*) \le \langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{x}^* \rangle \le \|\nabla f(\mathbf{x}^k)\|_* \|\mathbf{x}^k - \mathbf{x}^*\| \le R\|\nabla f(\mathbf{x}^k)\|_*.$$

- Combining the above with sufficient decrease property, $f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \ge \frac{1}{2L_f}\|\nabla f(\mathbf{x}^k)\|_*^2$, the result follows.

$O(1/k)$ rate of convergence of the non-Euclidean gradient method

For any $k \geq 1$,

$$f(\mathbf{x}^k) - f_{\rm opt} \leq \frac{2L_f R^2}{k}$$

Proof.

- Define $a_k = f(\mathbf{x}^k) - f_{\rm opt}$
- Then by previous lemma,

$$a_k - a_{k+1} \geq \frac{1}{C} a_k^2,$$

  where $C = 2L_f R^2$.
- We can thus deduce (why?) that $a_k \leq \frac{C}{k}$.

Non-Euclidean Gradient under the l_1-Norm

- $\mathbb{R}^n$ endowed with the l_1-norm.
- f be an L_f-smooth function w.r.t. the l_1-norm.

Non-Euclidean Gradient under the l_1-Norm

- **Initialization:** pick $\mathbf{x}^0 \in \mathbb{R}^n$.
- **General step:** for any $k = 0, 1, 2, \ldots$ execute the following steps:
  - set $i_k \in \operatorname{argmax}_i \left| \frac{\partial f(\mathbf{x}^k)}{\partial x_i} \right|$;
  - $\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{\|\nabla f(\mathbf{x}^k)\|_\infty}{L_f} \operatorname{sgn}\left(\frac{\partial f(\mathbf{x}^k)}{\partial x_{i_k}}\right) \mathbf{e}_{i_k}$.

Coordinate descent-type method

Example

Consider the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n}\left\{\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}+\mathbf{b}^T\mathbf{x}\right\},$$

- $\mathbf{A}\in\mathbb{S}^n_{++}$ and $\mathbf{b}\in\mathbb{R}^n$.
- The underlying space is $\mathbb{E}=\mathbb{R}^n$ endowed with the l_p-norm ($p\in[1,\infty]$).
- f is $L_f^{(p)}$-smooth with

$$L_f^{(p)}=\|\mathbf{A}\|_{p,q}=\max_{\mathbf{x}}\{\|\mathbf{A}\mathbf{x}\|_q:\|\mathbf{x}\|_p\leq 1\}$$

  with $q\in[1,\infty]$ satisfying $\frac{1}{p}+\frac{1}{q}=1$.

Two settings:

- $p=2$. In this case, since $\mathbf{A}$ is positive definite, $L_f^{(2)}=\|\mathbf{A}\|_{2,2}=\lambda_{\max}(\mathbf{A})$.
- $p=1$. Here $L_f^{(1)}=\|\mathbf{A}\|_{1,\infty}=\max_{i,j}|A_{i,j}|$.

Two Algorithms

Euclidean ($p = 2$):

Algorithm G2

- **Initialization:** pick $\mathbf{x}^0 \in \mathbb{R}^n$.
- **General step ($k \geq 0$):** $\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{1}{L_f^{(2)}}(\mathbf{A}\mathbf{x}^k + \mathbf{b})$.

Non-Euclidean ($p = 1$)

Algorithm G1

- **Initialization:** pick $\mathbf{x}^0 \in \mathbb{R}^n$.
- **General step ($k \geq 0$):**
  - pick $i_k \in \underset{i=1,2,\ldots,n}{\operatorname{argmax}} |\mathbf{A}_i \mathbf{x}^k + b_i|$, where $\mathbf{A}_i$ denotes ith row of $\mathbf{A}$.
  - update $\mathbf{x}_j^{k+1} = \begin{cases} \mathbf{x}_j^k, & j \neq i_k, \\ \mathbf{x}_{i_k}^k - \frac{1}{L_f^{(1)}}(\mathbf{A}_{i_k}\mathbf{x}^k + b_{i_k}), & j = i_k. \end{cases}$

• Algorithm G2 requires $O(n^2)$ operations per iteration, while algorithm G1 requires only $O(n)$.

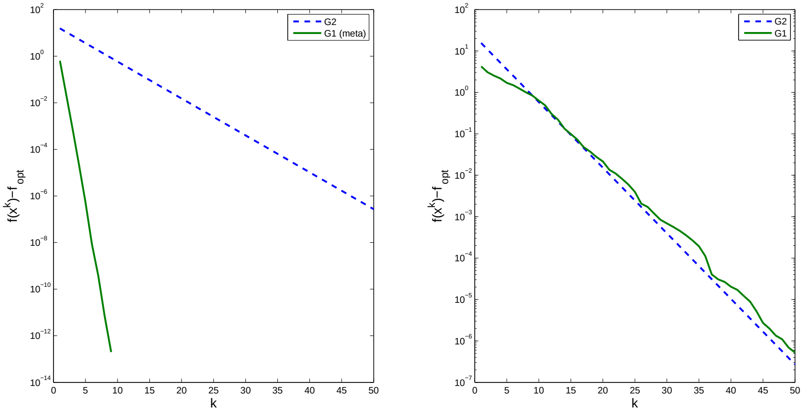
Example Contd.

- Set $\mathbf{A} = \mathbf{J} + 2\mathbf{I}$, where $\mathbf{J}$ is the matrix of all-ones.
- $\mathbf{A}$ is positive definite and $\lambda_{\max}(\mathbf{A}) = 2 + n$, $\max_{i,j} |A_{i,j}| = 3$.
- Therefore, as $\rho_f \equiv \frac{L_f^{(2)}}{L_f^{(1)}} = \frac{n+2}{3}$ gets larger, the Euclidean gradient method (Algorithm G2) should become more inferior to the non-Euclidean version (Algorithm G1).

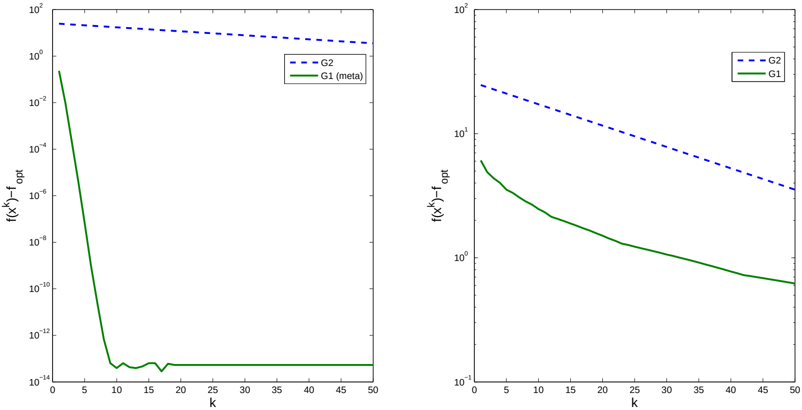
Numerical Example:

- $\mathbf{b} = 10\mathbf{e}_1, \mathbf{x}^0 = \mathbf{e}_n$.
- $n = 10/100(\rho_f = 4/34)$
- We count both iterations and "meta iterations" of G1.

$n = 10$



$n = 100$



Fast Proximal Gradient

- A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci. (2009).
- A. Beck and M. Teboulle, *Gradient-based algorithms with applications to signal-recovery problems*, In Convex optimization in signal processing and communications (2010)
- Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program. (2013)

FISTA (Fast Proximal Gradient Method)

- **The model:**

$$(P) \ \min_{\mathbf{x}\in\mathbb{E}} f(\mathbf{x}) + g(\mathbf{x})$$

- **Underlying Assumptions:**

(A) $g : \mathbb{E} \to (-\infty, \infty]$ is proper closed and convex.

(B) $f : \mathbb{E} \to \mathbb{R}$ is L_f-smooth and convex.

(C) The optimal set of (P) is nonempty and denoted by X^*. The optimal value of the problem is denoted by F_{opt}.

- **The Idea:** instead of making a step of the form

$$\mathbf{x}^{k+1} = \mathrm{prox}_{\frac{1}{L_k}g}\left(\mathbf{x}^k - \frac{1}{L_k}\nabla f(\mathbf{x}^k)\right)$$

we will consider a step of the form

$$\mathbf{x}^{k+1} = \mathrm{prox}_{\frac{1}{L_k}g}\left(\mathbf{y}^k - \frac{1}{L_k}\nabla f(\mathbf{y}^k)\right)$$

where $\mathbf{y}^k$ is a special linear combination of $\mathbf{x}^k, \mathbf{x}^{k-1}$

FISTA

FISTA
Input: $(f, g, \mathbf{x}^0)$, where f and g satisfy properties (A) and (B) and $\mathbf{x}^0 \in \mathbb{E}$.
Initialization: set $\mathbf{y}^0 = \mathbf{x}^0$ and $t_0 = 1$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick $L_k > 0$.

(b) set $\mathbf{x}^{k+1} = \mathrm{prox}_{\frac{1}{L_k} g}\left(\mathbf{y}^k - \frac{1}{L_k}\nabla f(\mathbf{y}^k)\right)$.

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

(d) compute $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.

- The dominant computational steps of the proximal gradient and FISTA methods are the same: one proximal computation and one gradient evaluation.

Stepsize Rules

- **constant.** $L_k = L_f$ for all k.
- **backtracking procedure B3.** The procedure requires two parameters (s, η), where $s > 0$ and $\eta > 1$. Define $L_{-1} = s$. At iteration k, L_k is set to be equal to L_{k-1}. Then, while

$$f(T_{L_k}(\mathbf{y}^k)) > f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k), T_{L_k}(\mathbf{y}^k) - \mathbf{y}^k \rangle + \frac{L_k}{2}\|T_{L_k}(\mathbf{y}^k) - \mathbf{y}^k\|^2,$$

we set $L_k := \eta L_k$. In other words, the stepsize is chosen as $L_k = L_{k-1}\eta^{i_k}$, where i_k is the smallest nonnegative integer for which

$$\begin{aligned} f(T_{L_{k-1}\eta^{i_k}}(\mathbf{y}^k)) \quad \leq \quad & f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k), T_{L_{k-1}\eta^{i_k}}(\mathbf{y}^k) - \mathbf{y}^k \rangle + \\ & \frac{L_k}{2}\|T_{L_{k-1}\eta^{i_k}}(\mathbf{y}^k) - \mathbf{y}^k\|^2. \end{aligned}$$

In both stepsize rules,

$$f(T_{L_k}(\mathbf{y}^k)) \leq f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k), T_{L_k}(\mathbf{y}^k) - \mathbf{y}^k \rangle + \frac{L_k}{2}\|T_{L_k}(\mathbf{y}^k) - \mathbf{y}^k\|^2.$$

Remarks

- $\beta L_f \le L_k \le \alpha L_f$, where

$$\alpha = \begin{cases} 1, & \text{constant,} \\ \max\left\{\eta, \frac{s}{L_f}\right\}, & \text{backtracking,} \end{cases} \quad \beta = \begin{cases} 1, & \text{constant,} \\ \frac{s}{L_f}, & \text{backtracking.} \end{cases}$$

- Easy to show by induction that $t_k \ge \frac{k+2}{2}$ for all $k \ge 0$.

$O(1/k^2)$ rate of convergence of FISTA

Theorem. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by FISTA with either a constant stepsize rule or the backtracking procedure B3. Then for any $\mathbf{x}^* \in X^*$ and $k \geq 1$,

$$F(\mathbf{x}^k) - F_{\rm opt} \leq \frac{2\alpha L_f \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{(k+1)^2},$$

where $\alpha = 1$ in the constant stepsize setting and $\alpha = \max\left\{\eta, \frac{s}{L_f}\right\}$ if the backtracking rule is employed.

Proof heavily based on the fundamental proximal gradient inequality.

Alternative Choice for t_k

- For the proof of the $O(1/k^2)$ rate, it is enough to require that $\{t_k\}_{k\geq 0}$ will satisfy
  (a) $t_k \geq \frac{k+2}{2}$;
  (b) $t_{k+1}^2 - t_{k+1} \leq t_k^2$.
- The choice $t_k = \frac{k+2}{2}$ also satisfies these two properties. (a) is obvious. (b) holds since

$$
\begin{aligned}
t_{k+1}^2 - t_{k+1} &= t_{k+1}(t_{k+1}-1) = \frac{k+3}{2}\cdot\frac{k+1}{2} = \frac{k^2+4k+3}{4} \\
&\leq \frac{k^2+4k+4}{4} = \frac{(k+2)^2}{4} = t_k^2.
\end{aligned}
$$

ISTA/FISTA

Consider the model

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

- $\lambda > 0$
- $f : \mathbb{R}^n \to \mathbb{R}$ convex and L_f-smooth.

Iterative Shrinkage/Thresholding Algorithm (ISTA):

$$\mathbf{x}^{k+1} = \mathcal{T}_{\lambda/L_f}\left(\mathbf{x}^k - \frac{1}{L_f}\nabla f(\mathbf{x}^k)\right).$$

Fast Iterative Shrinkage/Thresholding Algorithm (ISTA):

(a) $\mathbf{x}^{k+1} = \mathcal{T}_{\frac{\lambda}{L_f}}\left(\mathbf{y}^k - \frac{1}{L_f}\nabla f(\mathbf{y}^k)\right)$.

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

(c) $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{t_k-1}{t_{k+1}}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.

l_1-Regularized Least Squares

Consider the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1,$$

- $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$ and $\lambda > 0$.
- Fits (P) with $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$ and $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$.
- f is L_f-smooth with $L_f = \|\mathbf{A}^T\mathbf{A}\|_{2,2} = \lambda_{\max}(\mathbf{A}^T\mathbf{A})$.

ISTA: $\mathbf{x}^{k+1} = \mathcal{T}_{\frac{\lambda}{L_k}}\left(\mathbf{x}^k - \frac{1}{L_k}\mathbf{A}^T(\mathbf{A}\mathbf{x}^k-\mathbf{b})\right).$

FISTA:

(a) $\mathbf{x}^{k+1} = \mathcal{T}_{\frac{\lambda}{L_k}}\left(\mathbf{y}^k - \frac{1}{L_k}\mathbf{A}^T(\mathbf{A}\mathbf{y}^k-\mathbf{b})\right).$

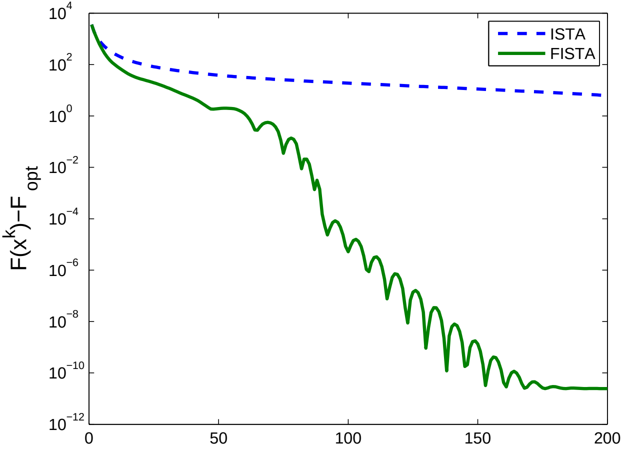
(b) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}.$

(c) $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{t_k-1}{t_{k+1}}\right)(\mathbf{x}^{k+1}-\mathbf{x}^k).$

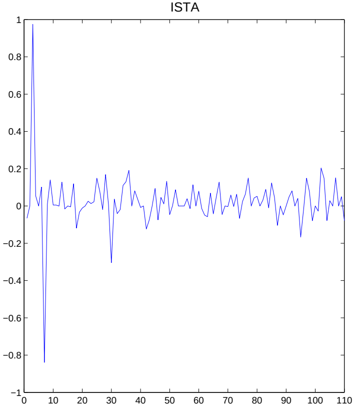
Numerical Example I

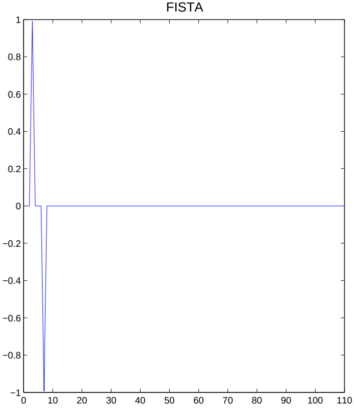
- test on regularized l_1-regularized least squares.
- $\lambda = 1$.
- $\mathbf{A} \in \mathbb{R}^{100\times 110}$. The components of $\mathbf{A}$ were independently generated using a standard normal distribution.
- the "true" vector is $\mathbf{x}_{\rm true} = \mathbf{e}_3 - \mathbf{e}_7$.
- $\mathbf{b} = \mathbf{A}\mathbf{x}_{\rm true}$.
- ran 200 iterations of ISTA and FISTA with $\mathbf{x}^0 = \mathbf{e}$.

Function Values



Solutions


ISTA
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
0 10 20 30 40 50 60 70 80 90 100 110


FISTA
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
0 10 20 30 40 50 60 70 80 90 100 110

Example 2: Wavelet-Based Image Deblurring

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 + \lambda\|\mathbf{x}\|_1$$

- image of size 512x512
- matrix **A** is dense (Gaussian blurring times inverse of two-stage Haar wavelet transform).
- all problems solved with fixed λ and Gaussian noise.

Deblurring of the Cameraman

original

blurred and noisy

1000 Iterations of ISTA versus 200 of FISTA

ISTA: **1000 Iterations**

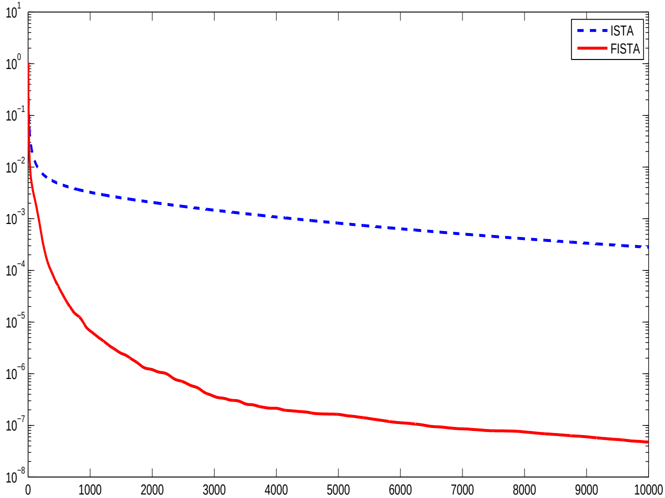
FISTA: **200 Iterations**

Original Versus Deblurring via FISTA

Original

FISTA:1000 Iterations

Function Values errors $F(\mathbf{x}^k) - F(\mathbf{x}^*)$



Weighted FISTA

- $\mathbb{E} = \mathbb{R}^n$
- The underlying assumption is that $\mathbb{E}$ is Euclidean.
- Assume that the endowed inner product is the **Q**-inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{Q} \mathbf{y},$$

  where $\mathbf{Q} \in \mathbb{S}^n_{++}$.
- $\nabla f(\mathbf{x}) = \mathbf{Q}^{-1} D_f(\mathbf{x})$, where

$$D_f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \frac{\partial f}{\partial x_2}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{pmatrix}.$$

- $L_f^{\mathbf{Q}}$ (Lipschitz constant of f w.r.t. the **Q**-norm):

$$\|\mathbf{Q}^{-1} D_f(\mathbf{x}) - \mathbf{Q}^{-1} D_f(\mathbf{y})\|_{\mathbf{Q}} \leq L_f^{\mathbf{Q}} \|\mathbf{x} - \mathbf{y}\|_{\mathbf{Q}} \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

Weighted FISTA

The general update rule for FISTA in this case will have the form

(a) $\mathbf{x}^{k+1} = \mathrm{prox}_{\frac{1}{L_f^{\mathbf{Q}}} g}\left(\mathbf{y}^k - \frac{1}{L_f}\mathbf{Q}^{-1} D_f(\mathbf{y}^k)\right)$.

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

(c) $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.

The prox operator in step (a) is computed in terms of the **Q**-norm:

$$\mathrm{prox}_h(\mathbf{x}) = \underset{\mathbf{u}\in\mathbb{R}^n}{\operatorname{argmin}}\left\{h(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|_{\mathbf{Q}}^2\right\}.$$

The convergence result will also be written in term of the **Q**-norm

$$F(\mathbf{x}^k) - F_{\mathrm{opt}} \le \frac{2\alpha L_f^{\mathbf{Q}}\|\mathbf{x}^0 - \mathbf{x}^*\|_{\mathbf{Q}}^2}{(k+1)^2}.$$

Restarting FISTA in the Strongly Convex Case

- Assume that f is σ-strongly convex for some $\sigma > 0$.
- The proximal gradient method attains an ε-optimal solution after an order of $O(\kappa \log(\frac{1}{\varepsilon}))$ iterations ($\kappa = \frac{L_f}{\sigma}$).
- A natural question is how the complexity result improves when using FISTA.
- Done by incorporating a restarting mechanism to FISTA – improves complexity result to $O(\sqrt{\kappa} \log(\frac{1}{\varepsilon}))$

Restarted FISTA
Initialization: pick $\mathbf{z}^{-1} \in \mathbb{E}$ and a positive integer N. Set $\mathbf{z}^0 = T_{L_f}(\mathbf{z}^{-1})$.
General step $(k \geq 0)$

- run N iterations of FISTA with constant stepsize $(L_k \equiv L_f)$ and input $(f, g, \mathbf{z}^k)$ and obtain a sequence $\{\mathbf{x}^n\}_{n=0}^N$;
- set $\mathbf{z}^{k+1} = \mathbf{x}^N$.

Restarted FISTA

Theorem [$O(\sqrt{\kappa}\log(\frac{1}{\varepsilon}))$ complexity of restarted FISTA] Suppose that that f is σ-strongly convex ($\sigma > 0$). Let $\{\mathbf{z}^k\}_{k\geq 0}$ be the sequence generated by the restarted FISTA method employed with $N = \lceil\sqrt{8\kappa} - 1\rceil$. Let R be an upper bound on $\|\mathbf{z}^{-1} - \mathbf{x}^*\|$. Then

(a) $F(\mathbf{z}^k) - F_{\rm opt} \leq \frac{L_f R^2}{2}\left(\frac{1}{2}\right)^k$;

(b) after k iterations of FISTA with k satisfying

$$k \geq \sqrt{8\kappa}\left(\frac{\log(\frac{1}{\varepsilon})}{\log(2)} + \frac{\log(L_f R^2)}{\log(2)}\right),$$

an ε-optimal solution is obtained at the end of last completed cycle:

$$F(\mathbf{z}^{\lfloor \frac{k}{N}\rfloor}) - F_{\rm opt} \leq \varepsilon.$$

Smoothing

- A. Beck and M. Teboulle, *Smoothing and first order methods: a unified framework.* SIAM J. Optim. (2012)
- Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program. (2005)

Smoothing

- It is known that in general smooth convex optimization problems can be solved with complexity $O(1/\varepsilon^2)$
- FISTA requires $O(1/\sqrt{\varepsilon})$ to obtain an ε-optimal solution of the composite model $f + g$.
- We will show how FISTA can be used to devise a method for more general nonsmooth convex problems in an improved complexity of $O(1/\varepsilon)$.

The model under consideration is

$$(P) \quad \min\{f(\mathbf{x}) + h(\mathbf{x}) + g(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}.$$

- f L_f-smooth and convex;
- g proper closed and convex and "proximable";
- h real-valued and convex (but not "proximable")

The Idea

$$(P) \quad \min\{f(\mathbf{x}) + h(\mathbf{x}) + g(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}.$$

- Solving (P) with FISTA with smooth/nosmooth parts $(f, g + h)$ is not practical.
- The idea will be to find a smooth approximation of h, say $\tilde{h}$ and solve the problem via FISTA with smooth and nonsmooth parts taken as $(f + \tilde{h}, g)$.
- This simple idea will be the basis for the improved $O(1/\varepsilon)$ complexity.
- Need to study in more details the notions of smooth approximations and smoothability.

Smooth Approximations and Smoothability

- Definition. A convex function $h : \mathbb{E} \to \mathbb{R}$ is called (α, β)-smoothable $(\alpha, \beta > 0)$ if for any $\mu > 0$ there exists a convex differentiable function $h_\mu : \mathbb{E} \to \mathbb{R}$ such that
  (a) $h_\mu(\mathbf{x}) \le h(\mathbf{x}) \le h_\mu(\mathbf{x}) + \beta\mu$ for all $\mathbf{x} \in \mathbb{E}$.
  (b) h_μ is $\frac{\alpha}{\mu}$-smooth.
- The function h_μ is called a $\frac{1}{\mu}$-smooth approximation of h with parameters (α, β).

Examples:

- $h(\mathbf{x}) = \|\mathbf{x}\|_2 (\mathbb{E} = \mathbb{R}^n)$. For any $\mu > 0$, $h_\mu(\mathbf{x}) \equiv \sqrt{\|\mathbf{x}\|_2^2 + \mu^2} - \mu$ is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(1, 1) \Rightarrow h$ is (1,1)-smoothable.
- $h(\mathbf{x}) = \max\{x_1, x_2, \ldots, x_n\} (\mathbb{E} = \mathbb{R}^n)$. For any $\mu > 0$, $h_\mu(\mathbf{x}) = \mu \log\left(\sum_{i=1}^n e^{x_i/\mu}\right) - \mu \log n$ is a smooth approximation of h with parameters $(1, \log n) \Rightarrow h$ is $(1, \log n)$-smoothable.

Calculus of Smooth Approximations

Theorem.

(a) Let $h^1, h^2 : \mathbb{E} \to \mathbb{R}$ be convex functions and let γ_1, γ_2 be nonnegative numbers. Suppose that for a given $\mu > 0$, h^i_μ is a $\frac{1}{\mu}$-smooth approximation of h^i with parameters (α_i, β_i) for $i = 1, 2$, then $\gamma_1 h^1_\mu + \gamma_2 h^2_\mu$ is a $\frac{1}{\mu}$-smooth approximation of $\gamma_1 h^1 + \gamma_2 h^2$ with parameters $(\gamma_1\alpha_1 + \gamma_2\alpha_2, \gamma_1\beta_1 + \gamma_2\beta_2)$.

(b) Let $\mathcal{A} : \mathbb{E} \to \mathbb{V}$ be a linear transformation between the Euclidean spaces $\mathbb{E}$ and $\mathbb{V}$. Let $h : \mathbb{V} \to \mathbb{R}$ be a convex function and define

$$q(\mathbf{x}) \equiv h(\mathcal{A}(\mathbf{x}) + \mathbf{b}),$$

where $\mathbf{b} \in \mathbb{V}$. Suppose that for a given $\mu > 0$, h_μ is a $\frac{1}{\mu}$-smooth approximation of h with parameters (α, β). Then the function $q_\mu(\mathbf{x}) \equiv h_\mu(\mathcal{A}(\mathbf{x}) + \mathbf{b})$ is a $\frac{1}{\mu}$-smooth approximation of q with parameters $(\alpha\|\mathcal{A}\|^2, \beta)$.

Proof: very easy...

Operations Preserving Smoothability

Corollary.

(a) Let $h^1, h^2 : \mathbb{E} \to \mathbb{R}$ be convex functions which are (α_1, β_1)- and (α_2, β_2)-smoothable respectively, and let γ_1, γ_2 be nonnegative numbers. Then $\gamma_1 h^1 + \gamma_2 h^2$ is a $(\gamma_1\alpha_1 + \gamma_2\alpha_2, \gamma_1\beta_1 + \gamma_2\beta_2)$-smoothable function.

(b) Let $\mathcal{A} : \mathbb{E} \to \mathbb{V}$ be a linear transformation between the Euclidean spaces $\mathbb{E}$ and $\mathbb{V}$. Let $h : \mathbb{V} \to \mathbb{R}$ be a convex (α, β)-smoothable function and define

$$q(\mathbf{x}) \equiv g(\mathcal{A}(\mathbf{x}) + \mathbf{b}),$$

where $\mathbf{b} \in \mathbb{V}$. Then q is an $(\alpha\|\mathcal{A}\|^2, \beta)$-smoothable function.

Smooth Approximation of Piecewise Affine Functions

- Let $q(\mathbf{x}) = \max_{i=1,\ldots,m}\{\mathbf{a}_i^T\mathbf{x} + b_i\}$, where $\mathbf{a}_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ for any $i = 1, 2, \ldots, m$.
- $q(\mathbf{x}) = g(\mathbf{A}\mathbf{x} + \mathbf{b})$, where $g(\mathbf{y}) = \max\{y_1, y_2, \ldots, y_m\}$, $\mathbf{A}$ is the matrix whose rows are $\mathbf{a}_1^T, \mathbf{a}_2^T, \ldots, \mathbf{a}_m^T$ and $\mathbf{b} = (b_1, b_2, \ldots, b_m)^T$.
- Let $\mu > 0$. $g_\mu(\mathbf{y}) = \mu \log\left(\sum_{i=1}^m e^{y_i/\mu}\right) - \mu \log m$ is a $\frac{1}{\mu}$-smooth approximation of g with parameters $(1, \log m)$.
- Therefore,

$$q_\mu(\mathbf{x}) \equiv g_\mu(\mathbf{A}\mathbf{x} + \mathbf{b}) = \mu \log\left(\sum_{i=1}^m e^{(\mathbf{a}_i^T\mathbf{x} + b_i)/\mu}\right) - \mu \log m$$

is a $\frac{1}{\mu}$-smooth approximation of q with parameters $(\|\mathbf{A}\|_{2,2}^2, \log m)$.

The Moreau Envelope

Definition. Given a proper closed convex function $f : \mathbb{E} \to (-\infty, \infty]$, and $\mu > 0$, the Moreau envelope of f is the function

$$M_f^{\mu}(\mathbf{x}) = \min_{\mathbf{u} \in \mathbb{E}} \left\{ f(\mathbf{u}) + \frac{1}{2\mu} \|\mathbf{x} - \mathbf{u}\|^2 \right\}.$$

- The parameter μ is called the smoothing parameter.
- By the first prox theorem the minimization problem defining the Moreau envelope has a unique solution, given by $\mathrm{prox}_{\mu f}(\mathbf{x})$. Therefore,

$$M_f^{\mu}(\mathbf{x}) = f(\mathrm{prox}_{\mu f}(\mathbf{x})) + \frac{1}{2\mu} \|\mathbf{x} - \mathrm{prox}_{\mu f}(\mathbf{x})\|^2.$$

Examples

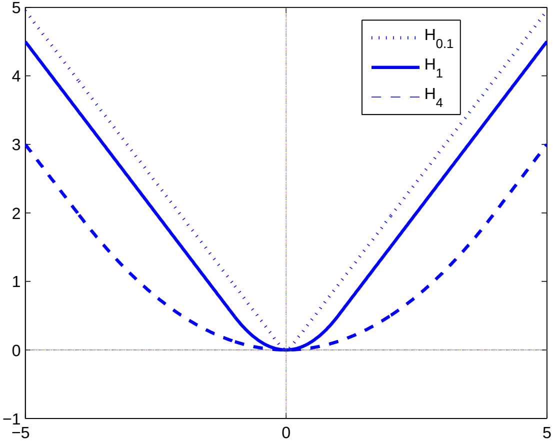
- **Indicators**. Suppose that $f = \delta_C$, where $C \subseteq \mathbb{E}$ is a nonempty closed and convex set. Then $\mathrm{prox}_f = P_C$ and
$$M_f^\mu(\mathbf{x}) = \delta_C(P_C(\mathbf{x})) + \frac{1}{2\mu}\|\mathbf{x} - P_C(\mathbf{x}))\|^2.$$
Therefore,
$$M_{\delta_C}^\mu = \frac{1}{2\mu} d_C^2.$$

- **Euclidean Norms** $f(\mathbf{x}) = \|\mathbf{x}\|$. Then for any $\mu > 0$ and $\mathbf{x} \in \mathbb{E}$,
$$\mathrm{prox}_{\mu f}(\mathbf{x}) = \left(1 - \frac{\mu}{\max\{\|\mathbf{x}\|, \mu\}}\right)\mathbf{x}.$$
Therefore,
$$M_f^\mu(\mathbf{x}) = \|\mathrm{prox}_{\mu f}(\mathbf{x})\| + \frac{1}{2\mu}\|\mathbf{x} - \mathrm{prox}_{\mu f}(\mathbf{x})\|^2 = \underbrace{\begin{cases} \frac{1}{2\mu}\|\mathbf{x}\|^2, & \|\mathbf{x}\| \le \mu, \\ \|\mathbf{x}\| - \frac{\mu}{2}, & \|\mathbf{x}\| > \mu, \end{cases}}_{H_\mu(\mathbf{x})}$$

H_μ - Huber function

Huber Function

H_μ gets smoother as μ increases.



Smoothability of the Moreau Envelope

Theorem. Let $f:\mathbb{E}\to(-\infty,\infty]$ be a proper closed and convex function. Let $\mu>0$. Then M_f^μ is $\frac{1}{\mu}$-smooth over $\mathbb{E}$ and

$$\nabla M_f^\mu(\mathbf{x})=\frac{1}{\mu}\left(\mathbf{x}-\mathrm{prox}_{\mu f}(\mathbf{x})\right).$$

Examples:

- **(smoothability of the squared distance)** Let $C\subseteq\mathbb{E}$ be a nonempty closed and convex set. Recall that $\frac{1}{2}d_C^2=M_{\delta_C}^1$. Then $\frac{1}{2}d_C^2$ is 1-smooth and
$$\nabla\left(1/2d_C^2\right)(\mathbf{x})=\mathbf{x}-\mathrm{prox}_{\delta_C}(\mathbf{x})=\mathbf{x}-P_C(\mathbf{x}).$$

- **(smoothability of Huber)** $H_\mu=M_f^\mu$, where $f(\mathbf{x})=\|\mathbf{x}\|$. Then H_μ is $\frac{1}{\mu}$-smooth and
$$\begin{aligned}\nabla H_\mu(\mathbf{x}) &= \frac{1}{\mu}\left(\mathbf{x}-\mathrm{prox}_{\mu f}(\mathbf{x})\right)=\frac{1}{\mu}\left(\mathbf{x}-\left(1-\frac{\mu}{\max\{\|\mathbf{x}\|,\mu\}}\right)\mathbf{x}\right)\\ &= \begin{cases}\frac{1}{\mu}\mathbf{x}, & \|\mathbf{x}\|\le\mu,\\ \frac{\mathbf{x}}{\|\mathbf{x}\|}, & \|\mathbf{x}\|>\mu,\end{cases}\end{aligned}$$

Smoothability of Lipschitz Convex Functions

Theorem. Let $h : \mathbb{E} \to \mathbb{R}$ be a convex function satisfying

$$|h(\mathbf{x}) - h(\mathbf{y})| \le \ell_h \|\mathbf{x} - \mathbf{y}\| \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{E}.$$

Then $\mu > 0$ M_h^μ is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(1, \frac{\ell_h^2}{2})$.

Corollary. Let $h : \mathbb{E} \to \mathbb{R}$ be convex and Lipschitz with constant ℓ_h. Then h is $(1, \frac{\ell_h^2}{2})$-smoothable.

Examples:

- **(smooth approximation of the l_2-norm)** Let $h(\mathbf{x}) = \|\mathbf{x}\|_2$ (over $\mathbb{R}^n$). Then h is convex and Lipschitz with constant $\ell_h = 1$. Therefore,

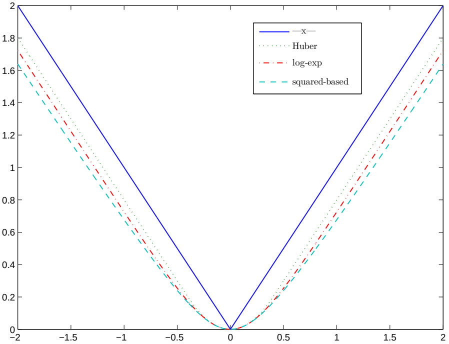
$$M_h^\mu(\mathbf{x}) = H_\mu(\mathbf{x}) = \begin{cases} \frac{1}{2\mu}\|\mathbf{x}\|_2^2, & \|\mathbf{x}\|_2 \le \mu, \\ \|\mathbf{x}\|_2 - \frac{\mu}{2}, & \|\mathbf{x}\|_2 > \mu. \end{cases}$$

  is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(1, \frac{1}{2})$.
- **(smooth approximation of the l_1-norm)** Let $h(\mathbf{x}) = \|\mathbf{x}\|_1$ Then h is convex and Lipschitz with constant $\ell_h = \sqrt{n}$. Hence, $M_h^\mu(\mathbf{x}) = \sum_{i=1}^n H_\mu(x_i)$ is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(1, \frac{n}{2})$.

Smooth Approximations of the Absolute Value Function

Three possible smooth approximations of $h(x) = |x|$

- $h_\mu^1(x) = \sqrt{x^2 + \mu^2} - \mu$, $(\alpha, \beta) = (1, 1)$.
- $h_\mu^2(x) = \mu \log(e^{x/\mu} + e^{-x/\mu}) - \mu \log 2$, $(\alpha, \beta) = (1, \log 2)$.
- $h_\mu^3(x) = H_\mu(x)$, $(\alpha, \beta) = (1, \frac{1}{2})$.



Back to Algorithms - Model and Assumptions

Main model:

$$(P) \quad \min_{\mathbf{x}\in\mathbb{E}} \{H(\mathbf{x}) \equiv f(\mathbf{x}) + h(\mathbf{x}) + g(\mathbf{x})\}$$

.

(A) $f : \mathbb{E} \to \mathbb{R}$ is L_f-smooth ($L_f > 0$).

(B) $h : \mathbb{E} \to \mathbb{R}$ is (α, β)-smoothable ($\alpha, \beta > 0$). For any $\mu > 0$, h_μ denotes a $\frac{1}{\mu}$-smooth approximation of h with parameters (α, β).

(C) $g : \mathbb{E} \to (-\infty, \infty]$ is proper closed and convex.

(D) H has bounded level sets. Specifically, for any $\delta > 0$, there exists $R_\delta > 0$ such that

$$\|\mathbf{x}\| \leq R_\delta \text{ for any } \mathbf{x} \text{ satisfying } H(\mathbf{x}) \leq \delta.$$

(E) The optimal set of (P) is nonempty and denoted by X^*. The optimal value of the problem is denoted by $H_{\rm opt}$.

The S-FISTA Method

- The idea is to consider the following smoothed version of (P):

$$(P_\mu) \quad \min_{\mathbf{x}\in\mathbb{E}}\{H_\mu(\mathbf{x}) \equiv \underbrace{f(\mathbf{x}) + h_\mu(\mathbf{x})}_{F_\mu(\mathbf{x})} + g(\mathbf{x})\},$$

  for some $\mu > 0$, and solve it using FISTA with constant stepsize.
- A Lipschitz constant of ∇F_μ is $L_f + \frac{\alpha}{\mu}$; the stepsize is taken as $\frac{1}{L_f+\frac{\alpha}{\mu}}$.

S-FISTA
Input: $\mathbf{x}^0 \in \mathrm{dom}(g)$, $\mu > 0$.
Initialization: set $\mathbf{y}^0 = \mathbf{x}^0$, $t_0 = 1$; construct h_μ – a $\frac{1}{\mu}$-smooth approximation of h with parameters (α, β); set $F_\mu = f + h_\mu$, $\tilde{L} = L_f + \frac{\alpha}{\mu}$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) $\mathbf{x}^{k+1} = \mathrm{prox}_{\frac{1}{\tilde{L}}g}\left(\mathbf{y}^k - \frac{1}{\tilde{L}}\nabla F_\mu(\mathbf{y}^k)\right)$;

(b) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;

(c) $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{t_k-1}{t_{k+1}}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.

$O(1/\varepsilon)$ complexity of S-FISTA

Theorem. Let $\varepsilon \in (0, \bar{\varepsilon})$ for some fixed $\bar{\varepsilon}$. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by S-FISTA with smoothing parameter

$$\mu = \sqrt{\frac{\alpha}{\beta}} \frac{\varepsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta + L_f \varepsilon}}.$$

Then for any k satisfying

$$k \geq 2\sqrt{2\alpha\beta\Gamma}\frac{1}{\varepsilon} + \sqrt{2L_f\Gamma}\frac{1}{\sqrt{\varepsilon}},$$

where $\Gamma = (R_{H(\mathbf{x}^0)+\frac{\bar{\varepsilon}}{2}} + \|\mathbf{x}^0\|)^2$, it holds that $H(\mathbf{x}^k) - H_{\rm opt} \leq \varepsilon$.

Minimization of "Proximable" Functions

Consider the problem

$$(P_1) \quad \min_{\mathbf{x}\in\mathbb{E}}\{h(\mathbf{x}) : \mathbf{x} \in C\},$$

- C is a nonempty closed and convex set.
- $h : \mathbb{E} \to \mathbb{R}$ is convex function Lipschitz with constant ℓ_h.
- Fits model (P) with $f = 0$ and $g = \delta_C$.
- $h_\mu = M_h^\mu$ is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(\alpha, \beta) = (1, \frac{\ell_h^2}{2})$.
- $\nabla M_h^\mu(\mathbf{x}) = \frac{1}{\mu}(\mathbf{x} - \mathrm{prox}_{\mu h}(\mathbf{x}))$.
- After employing $O(1/\varepsilon)$ iterations of the the S-FISTA method with

$$\mu = \sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta + L_f\varepsilon}} = \sqrt{\frac{\alpha}{\beta}}\frac{\varepsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta}} = \frac{\varepsilon}{2\beta} = \frac{\varepsilon}{\ell_h^2},$$

  an ε-optimal solution will be achieved.
- The stepsize is $\frac{1}{\tilde{L}}$, where $\tilde{L} = \frac{\alpha}{\mu} = \frac{1}{\mu}$.

S-FISTA for Solving (P_1)

- The general step of the S-FISTA method is

$$\begin{aligned}
\mathbf{x}^{k+1} &= \operatorname{prox}_{\frac{1}{\tilde{L}}g}\left(\mathbf{y}^k - \frac{1}{\tilde{L}}\nabla F_\mu(\mathbf{y}^k)\right) = P_C\left(\mathbf{y}^k - \frac{1}{\tilde{L}\mu}(\mathbf{y}^k - \operatorname{prox}_{\mu h}(\mathbf{y}^k))\right) \\
&= P_C(\operatorname{prox}_{\mu h}(\mathbf{y}^k)).
\end{aligned}$$

S-FISTA for solving (P_1)

Initialization: set $\mathbf{y}^0 = \mathbf{x}^0 \in C$, $t_0 = 1$; set $\mu = \frac{\varepsilon}{\ell_h^2}$ and $\tilde{L} = \frac{\ell_h^2}{\varepsilon}$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) $\mathbf{x}^{k+1} = P_C(\operatorname{prox}_{\mu h}(\mathbf{y}^k))$;

(b) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;

(c) $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.

Block Proximal Gradient Methods

- A. Beck and L. Tetruashvili. *On the convergence of block coordinate descent type methods*, SIAM J. Optim. (2013)
- M. Hong, X. Wang, M. Razaviyayn, and Z. Q Luo, *Iteration complexity analysis of block coordinate descent methods*, Arxiv.
- Q. Lin, Z. Lu, and L. Xiao, *An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization*, SIAM J. Optim., (2015)
- R. Shefi and M. Teboulle, *On the rate of convergence of the proximal alternating linearized minimization algorithm for convex problems*, EURO J. Comput. Optim. (2016)

Block Proximal Gradient Methods

The Model

$$(P) \qquad \min_{\mathbf{x}_1 \in \mathbb{E}_1, \mathbf{x}_2 \in \mathbb{E}_2, \ldots, \mathbf{x}_p \in \mathbb{E}_p} \left\{ F(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) = f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) + \sum_{j=1}^{p} g_j(\mathbf{x}_j) \right\},$$

Setting and Notation

- $\mathbb{E}_1, \mathbb{E}_2, \ldots, \mathbb{E}_p$ are Euclidean spaces.
- $\mathbb{E} = \mathbb{E}_1 \times \mathbb{E}_2 \times \cdots \times \mathbb{E}_p$. We use the notation that a vector $\mathbf{x} \in \mathbb{E}$ can be written as $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$.
- The product space is also Euclidean with endowed norm $\|(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p)\|_{\mathbb{E}} = \sqrt{\sum_{i=1}^{p} \|\mathbf{u}_i\|_{\mathbb{E}_i}^2}$.
- $g : \mathbb{E} \to (-\infty, \infty]$ is defined by $g(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) \equiv \sum_{i=1}^{p} g_i(\mathbf{x}_i)$. (P) can thus be simply written as $\min_{\mathbf{x} \in \mathbb{E}} f(\mathbf{x}) + g(\mathbf{x})$
- The gradient w.r.t. the ith block ($i \in \{1, 2, \ldots, p\}$) is denoted by $\nabla_i f$ $\nabla f(\mathbf{x}) = (\nabla_1 f(\mathbf{x}), \nabla_2 f(\mathbf{x}), \ldots, \nabla_p f(\mathbf{x}))$.
- For any $i \in \{1, 2, \ldots, p\}$ we define $\mathcal{U}_i : \mathbb{E}_i \to \mathbb{E}$ to be the linear transformation given by $\mathcal{U}_i(\mathbf{d}) = (\mathbf{0}, \ldots, \mathbf{0}, \underbrace{\mathbf{d}}_{i\text{th block}}, \mathbf{0}, \ldots, \mathbf{0}), \quad \mathbf{d} \in \mathbb{E}_i$.

Underlying Assumption

(A) $g_i : \mathbb{E}_i \to (-\infty, \infty]$ is proper closed and convex for any $i \in \{1, 2, \ldots, p\}$.

(B) $f : \mathbb{E} \to \mathbb{R}$ is L_f-smooth and convex.

(C) There exist $L_1, L_2, \ldots, L_p > 0$ such that for any $i \in \{1, 2, \ldots, p\}$ it holds that

$$\|\nabla_i f(\mathbf{x}) - \nabla_i f(\mathbf{x} + \mathcal{U}_i(\mathbf{d}))\| \leq L_i \|\mathbf{d}\|$$

for all $\mathbf{x} \in \mathbb{E}$ and $\mathbf{d} \in \mathbb{E}_i$.

(D) The optimal set of problem (P) is nonempty and denoted by X^*. The optimal value is denoted by $F_{\rm opt}$.

The Block Proximal Gradient Method

The Block Proximal Gradient Method

Initialization. pick $\mathbf{x}^0 = (\mathbf{x}_1^0, \mathbf{x}_2^0, \ldots, \mathbf{x}_p^0) \in \mathrm{int}(\mathrm{dom}(f))$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick $i_k \in \{1, 2, \ldots, p\}$;

(b) $$\mathbf{x}_j^{k+1} = \begin{cases} \mathrm{prox}_{\frac{1}{L_{i_k}} g_{i_k}}\left(\mathbf{x}_{i_k} - \frac{1}{L_{i_k}} \nabla_{i_k} f(\mathbf{x}^k)\right), & j = i_k, \\ \mathbf{x}_j^k, & j \neq i_k. \end{cases}$$

Index selection strategies:

- **cyclic.** $i_k = (k \mod p) + 1$.
  Cyclic Block Proximal Gradient (CBPG)
- **randomized.** i_k is randomly picked from $\{1, 2, \ldots, p\}$ by a uniform distribution.
  Randomized Block Proximal Gradient (RBPG)

$O(1/k)$ Rate of CBPG

Theorem. Suppose that Assumptions (A-D) hold as well as

(E) For any $\alpha > 0$, there exists $R_\alpha > 0$ such that

$$\max_{\mathbf{x},\mathbf{x}^* \in \mathbb{E}} \{\|\mathbf{x} - \mathbf{x}^*\| : F(\mathbf{x}) \le \alpha, \mathbf{x}^* \in X^*\} \le R_\alpha.$$

Let $\{\mathbf{x}^k\}_{k \ge 0}$ be the sequence generated by the CBPG method. For any $k \ge 2$:

$$F(\mathbf{x}^{pk}) - F_{\rm opt} \le \max\left\{ \left(\frac{1}{2}\right)^{(k-1)/2} (F(\mathbf{x}^0) - F_{\rm opt}), \frac{8p(L_f + L_{\max})^2 R^2}{L_{\min}(k-1)} \right\},$$

where $L_{\min} = \min_{i=1,2,\ldots,p} L_i$, $L_{\max} = \max_{i=1,2,\ldots,p} L_i$ and $R = R_{F(\mathbf{x}^0)}$.

$O(1/k)$ Rate of RBPG

Theorem. Suppose that Assumption (A)-(D) hold. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the RBPG method. Let $\mathbf{x}^* \in X^*$. Then for any $k \geq 0$,

$$\mathsf{E}_{\xi_k}(F(\mathbf{x}^{k+1})) - F_{\mathrm{opt}} \leq \frac{p}{p+k+1}\left(\frac{1}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|_L^2 + F(\mathbf{x}^0) - F_{\mathrm{opt}}\right).$$

Here

$$\|\mathbf{v}\|_L^2 \equiv \sqrt{\sum_{i=1}^{p} L_i \|\mathbf{v}_i\|^2}$$

Dual-Based Proximal Gradient Methods

- A. Beck and M. Teboulle, *A fast dual proximal gradient algorithm for convex minimization and applications*, Oper. Res. Lett. (2014)
- A. Beck and M. Teboulle. *Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems*, IEEE Trans. Image Process. (2009)
- A. Beck, L. Tetruashvili, Y. Vaisbourd, and A. Shemtov, *Rate of convergence analysis of dual-based variables decomposition methods for strongly convex problems*, (2016)
- A. Chambolle, *An algorithm for total variation minimization and applications*, J. Math. Imaging Vision (2004)
- P. Tseng, *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. SIAM J. Control Optim.*, (1991)

The Main Model

Main Model:

$$(P) \quad f_{\rm opt} = \min_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}) + g(\mathcal{A}(\mathbf{x})) \right\},$$

Underlying Assumptions:

(A) $f : \mathbb{E} \to (-\infty, +\infty]$ is proper closed and σ-strongly convex ($\sigma > 0$).

(B) $g : \mathbb{V} \to (-\infty, +\infty]$ is proper closed and convex.

(C) $\mathcal{A} : \mathbb{E} \to \mathbb{V}$ is a linear transformation.

(D) there exists $\hat{\mathbf{x}} \in \text{ri}(\text{dom}(f))$ and $\hat{\mathbf{z}} \in \text{ri}(\text{dom}(g))$ such that $\mathcal{A}(\hat{\mathbf{x}}) = \hat{\mathbf{z}}$.

Existence and uniqueness of optimal solution: under the above assumptions, the objective function is proper closed and strongly convex, and hence there exists a unique optimal solution, which will be denoted by $\mathbf{x}^*$.

Example 1: Orthogonal Projection onto a Polyhedral set

- Let
$$S = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{b}\},$$
where $\mathbf{A} \in \mathbb{R}^{p \times n}$, $\mathbf{b} \in \mathbb{R}^p$. Assume that $S \neq \emptyset$.
- Let $\mathbf{d} \in \mathbb{R}^n$. The orthogonal projection of $\mathbf{d}$ onto S is the unique optimal solution of
$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|^2 : \mathbf{A}\mathbf{x} \leq \mathbf{b} \right\}.$$
- Fits model (P) with $\mathbb{E} = \mathbb{R}^n, \mathbb{V} = \mathbb{R}^p$, $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|^2$,
$$g(\mathbf{z}) = \delta_{\text{Box}[-\infty\mathbf{e},\mathbf{b}]}(\mathbf{z}) = \begin{cases} \mathbf{0}, & \mathbf{z} \leq \mathbf{b}, \\ \infty, & \text{else.} \end{cases}$$
and $\mathcal{A}(\mathbf{x}) \equiv \mathbf{A}\mathbf{x}$.
- $\sigma = 1$

Example 2: One-Dimensional Total Variation Denoising

- **Denoising problem:**

$$\min_{\mathbf{x}\in\mathbb{E}} \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|^2 + R(\mathcal{A}(\mathbf{x})).$$

  - $\mathbf{d}\in\mathbb{E}$ - noisy and known signal
  - $\mathcal{A}:\mathbb{E}\to\mathbb{V}$ - linear transformation.
  - $R:\mathbb{V}\to\mathbb{R}_+$ - regularizing function measuring the magnitude of its argument.
- One-dimensional total variation denoising problem,
$\mathbb{E}=\mathbb{R}^n, \mathbb{V}=\mathbb{R}^{n-1}, \mathcal{A}(\mathbf{x})=\mathbf{D}\mathbf{x}, R(\mathbf{z})=\lambda\|\mathbf{z}\|_1(\lambda>0)$, $\mathbf{D}$ defined by
$\mathbf{D}\mathbf{x}=(x_1-x_2, x_2-x_3, \ldots, x_{n-1}-x_n)^T$

$$(P_1) \quad \min_{\mathbf{x}\in\mathbb{R}^n} \left\{\frac{1}{2}\|\mathbf{x}-\mathbf{d}\|_2^2 + \lambda\|\mathbf{D}\mathbf{x}\|_1\right\}.$$

- More explicitly: $\min_{\mathbf{x}\in\mathbb{E}} \left\{\frac{1}{2}\|\mathbf{x}-\mathbf{d}\|_2^2 + \lambda\sum_{i=1}^{n-1}|x_i-x_{i+1}|\right\}.$
- The function $\mathbf{x}\mapsto\|\mathbf{D}\mathbf{x}\|_1$ is a one-dimensional total variation function.
- Fits model (P) with
$\mathbb{E}=\mathbb{R}^n, \mathbb{V}=\mathbb{R}^{n-1}, f(\mathbf{x})=\frac{1}{2}\|\mathbf{x}-\mathbf{d}\|^2(\sigma=1), g(\mathbf{y})=\lambda\|\mathbf{y}\|_1,\ \mathcal{A}(\mathbf{x})\equiv\mathbf{D}\mathbf{x}$

The Dual Problem

- (P) is the same as $\min_{\mathbf{x},\mathbf{z}}\{f(\mathbf{x}) + g(\mathbf{z}) : \mathcal{A}(\mathbf{x}) - \mathbf{z} = \mathbf{0}\}$
- Lagrangian:
  $L(\mathbf{x}, \mathbf{z}; \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{z}) - \langle \mathbf{y}, \mathcal{A}(\mathbf{x}) - \mathbf{z}\rangle = f(\mathbf{x}) + g(\mathbf{z}) - \langle \mathcal{A}^T(\mathbf{y}), \mathbf{x}\rangle + \langle \mathbf{y}, \mathbf{z}\rangle.$
- Minimizing the Lagrangian w.r.t. **x** and **z**, we obtain the dual problem

$$\boxed{(D) \quad q_{\rm opt} = \max_{\mathbf{y}\in\mathbb{V}} \left\{ q(\mathbf{y}) \equiv -f^*(\mathcal{A}^T(\mathbf{y})) - g^*(-\mathbf{y}) \right\}.}$$

Theorem [strong duality of the pair (P),(D)] $f_{\rm opt} = q_{\rm opt}$ and the dual problem (D) attains an optimal solution.

The dual problem in minimization form:

$$\boxed{(D') \quad \min_{\mathbf{y}\in\mathbb{V}}\{F(\mathbf{y}) + G(\mathbf{y})\}}$$

$$\begin{aligned} F(\mathbf{y}) &\equiv f^*(\mathcal{A}^T(\mathbf{y})), \\ G(\mathbf{y}) &\equiv g^*(-\mathbf{y}). \end{aligned}$$

Rockafellar-Wets Theorem

Theorem [Rockafellar-Wets] Let $\sigma > 0$. Then

(a) If $f : \mathbb{E} \to \mathbb{R}$ is a $\frac{1}{\sigma}$-smooth convex function, then f^* is σ-strongly convex.

(b) If $f : \mathbb{E} \to (-\infty, \infty]$ is a proper closed σ-strongly convex function, then $f^* : \mathbb{E} \to \mathbb{R}$ is $\frac{1}{\sigma}$-smooth.

The Dual Problem

$$\boxed{(D') \quad \min_{\mathbf{y} \in \mathbb{V}} \{F(\mathbf{y}) + G(\mathbf{y})\}}$$

Properties of F and G:

(a) $F : \mathbb{V} \to \mathbb{R}$ is convex and L_F-smooth with $L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$;

(b) $G : \mathbb{V} \to (-\infty, \infty]$ is proper closed and convex.

Dual Proximal Gradient

Dual Proximal Gradient = Proximal Gradient on (D')

Dual Proximal Gradient – dual representation

- **Initialization:** pick $\mathbf{y}^0 \in \mathbb{V}$ and $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$.
- **General step** $(k \geq 0)$**:**

$$\mathbf{y}^{k+1} = \operatorname{prox}_{\frac{1}{L}G}\left(\mathbf{y}^k - \frac{1}{L}\nabla F(\mathbf{y}^k)\right).$$

Theorem [rate of convergence of the dual objective function] Let $\{\mathbf{y}^k\}_{k\geq 0}$ be the sequence generated by the DPG method with $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$. Then for any dual optimal solution $\mathbf{y}^*$ $k \geq 1$,

$$q_{\text{opt}} - q(\mathbf{y}^k) \leq \frac{L\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{2k}.$$

Constructing a Primal Representation–Technical Lemma

Lemma. Let $F(\mathbf{y}) = f^*(\mathcal{A}^T(\mathbf{y}) + \mathbf{b}), G(\mathbf{y}) = g^*(-\mathbf{y})$, where f, g and $\mathcal{A}$ satisfy properties (A),(B) and (C) and $\mathbf{b} \in \mathbb{E}$. Then for any $\mathbf{y}, \mathbf{v} \in \mathbb{V}$ and $L > 0$ the relation

$$\mathbf{y} = \mathrm{prox}_{\frac{1}{L}G}\left(\mathbf{v} - \frac{1}{L}\nabla F(\mathbf{v})\right) \tag{9}$$

holds if and only if

$$\mathbf{y} = \mathbf{v} - \frac{1}{L}\mathcal{A}(\tilde{\mathbf{x}}) + \frac{1}{L}\mathrm{prox}_{Lg}(\mathcal{A}(\tilde{\mathbf{x}}) - L\mathbf{v}),$$

where

$$\tilde{\mathbf{x}} = \operatorname*{argmax}_{\mathbf{x}} \left\{\langle \mathbf{x}, \mathcal{A}^T(\mathbf{v}) + \mathbf{b}\rangle - f(\mathbf{x})\right\}.$$

Dual Proximal Gradient - Primal Representation

The Dual Proximal Gradient (DPG) Method – primal representation

Initialization: pick $\mathbf{y}^0 \in \mathbb{V}$, and $L \geq \frac{\|\mathcal{A}\|^2}{\sigma}$.

General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) set $\mathbf{x}^k = \underset{\mathbf{x}}{\operatorname{argmax}} \left\{ \langle \mathbf{x}, \mathcal{A}^T(\mathbf{y}^k) \rangle - f(\mathbf{x}) \right\}$;

(b) set $\mathbf{y}^{k+1} = \mathbf{y}^k - \frac{1}{L}\mathcal{A}(\mathbf{x}^k) + \frac{1}{L}\mathrm{prox}_{Lg}(\mathcal{A}(\mathbf{x}^k) - L\mathbf{y}^k)$.

- The sequence $\{\mathbf{x}^k\}_{k \geq 0}$ generated by the method will be called "the primal sequence", although its elements are not necessarily feasible.

The Primal-Dual Relation

Obtaining a rate of the primal sequence is done using the following result.

Lemma [primal-dual relation] Let $\bar{\mathbf{y}} \in \text{dom}(G)$, and let

$$\bar{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{E}}{\text{argmax}} \left\{ \langle \mathbf{x}, \mathcal{A}^T(\bar{\mathbf{y}}) \rangle - f(\mathbf{x}) \right\}.$$

Then

$$\|\bar{\mathbf{x}} - \mathbf{x}^*\|^2 \leq \frac{2}{\sigma}(q_{\rm opt} - q(\bar{\mathbf{y}})).$$

$O(1/k)$ Rate of the Primal Sequence Generated by DPG

Theorem. Let $\{\mathbf{x}^k\}_{k\geq 0}$ and $\{\mathbf{y}^k\}_{k\geq 0}$ be the primal and dual sequences generated by the DPG method with $L \geq L_F$. Then for any optimal dual solution $\mathbf{y}^*$ and $k \geq 1$,

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{L\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{\sigma k}.$$

Proof.

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{2}{\sigma}(q_{\mathrm{opt}} - q(\mathbf{y}^k)) \leq \frac{2}{\sigma}\frac{L\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{2k},$$

Fast Dual Proximal Gradient (FDPG)

Fast Dual Proximal Gradient = FISTA on (D')

Fast Dual Proximal Gradient (FDPG) - dual representation

- **Initialization:** $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}, \mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{E}, t_0 = 1$.
- **General Step ($k \geq 0$):**
  (a) $\mathbf{y}^{k+1} = \mathrm{prox}_{\frac{1}{L}G}\left(\mathbf{w}^k - \frac{1}{L}\nabla F(\mathbf{w}^k)\right)$;
  (b) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
  (c) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.

Theorem [rate of convergence of the dual objective function] Let $\{\mathbf{y}^k\}_{k\geq 0}$ be the sequence generated by the FDPG method with $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$. Then for any dual optimal solution $\mathbf{y}^*$ of and $k \geq 1$,

$$q_{\mathrm{opt}} - q(\mathbf{y}^k) \leq \frac{2L\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{(k+1)^2}.$$

Fast Dual Proximal Gradient - Primal Representation

The Fast Dual Proximal Gradient (FDPG) Method - primal representation

Initialization: $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}, \mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{V}, t_0 = 1$.
General step $(k \geq 0)$**:**

(a) $\mathbf{u}^k = \underset{\mathbf{u}}{\operatorname{argmax}} \left\{ \langle \mathbf{u}, \mathcal{A}^T(\mathbf{w}^k) \rangle - f(\mathbf{u}) \right\}.$

(b) $\mathbf{y}^{k+1} = \mathbf{w}^k - \frac{1}{L}\mathcal{A}(\mathbf{u}^k) + \frac{1}{L}\mathrm{prox}_{Lg}(\mathcal{A}(\mathbf{u}^k) - L\mathbf{w}^k)$

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

(d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k).$

$O(1/k^2)$ Rate of the Primal Sequence Generated by FDPG

Theorem Let $\{\mathbf{x}^k\}_{k\geq 0}$ and $\{\mathbf{y}^k\}_{k\geq 0}$ be the primal and dual sequences generated by the FDPG method with $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$. Then for any optimal dual solution $\mathbf{y}^*$ and $k \geq 1$,

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{4L\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{\sigma(k+1)^2}.$$

Proof.

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{2}{\sigma}(q_{\rm opt} - q(\mathbf{y}^k)) \leq \frac{2}{\sigma} \cdot \frac{2L\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{(k+1)^2}.$$

Example 1: Orthogonal Projection onto a Polyhedral set

$$(P_1) \quad \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|^2 : \mathbf{A}\mathbf{x} \le \mathbf{b} \right\}.$$

- Fits model (P) with $\mathbb{E} = \mathbb{R}^n, \mathbb{V} = \mathbb{R}^p$, $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|^2$,

$$g(\mathbf{z}) = \delta_{\mathrm{Box}[-\infty \mathbf{e}, \mathbf{b}]}(\mathbf{z}) = \begin{cases} \mathbf{0}, & \mathbf{z} \le \mathbf{b}, \\ \infty, & \text{else}. \end{cases}$$

  and $\mathcal{A}(\mathbf{x}) \equiv \mathbf{A}\mathbf{x}$.
- $\sigma = 1$
- $\operatorname{argmax}_{\mathbf{x}}\{\langle \mathbf{v}, \mathbf{x}\rangle - f(\mathbf{x})\} = \mathbf{v} + \mathbf{d}$ for any $\mathbf{v} \in \mathbb{R}^n$;
- $\|\mathcal{A}\| = \|\mathbf{A}\|_{2,2}$;
- $\mathcal{A}^T(\mathbf{y}) = \mathbf{A}^T\mathbf{y}$ for any $\mathbf{y} \in \mathbb{R}^p$;
- $\mathrm{prox}_{Lg}(\mathbf{z}) = P_{\mathrm{Box}[-\infty \mathbf{e}, \mathbf{b}]}(\mathbf{z}) = \min\{\mathbf{z}, \mathbf{b}\}$.

DPG and FDPG for solving (P_1)

Algorithm 1 [DPG for solving (P_1)]

- **Initialization:** $L \geq \|\mathbf{A}\|_{2,2}^2, \mathbf{y}^0 \in \mathbb{R}^p$.
- **General Step** ($k \geq 0$):
  (a) $\mathbf{x}^k = \mathbf{A}^T\mathbf{y}^k + \mathbf{d}$;
  (b) $\mathbf{y}^{k+1} = \mathbf{y}^k - \frac{1}{L}\mathbf{A}\mathbf{x}^k + \frac{1}{L}\min\{\mathbf{A}\mathbf{x}^k - L\mathbf{y}^k, \mathbf{b}\}$.

Algorithm 2 [FDPG for solving (P_1)]

- **Initialization:** $L \geq \|\mathbf{A}\|_{2,2}^2, \mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{R}^p, t_0 = 1$.
- **General Step** ($k \geq 0$):
  (a) $\mathbf{u}^k = \mathbf{A}^T\mathbf{w}^k + \mathbf{d}$;
  (b) $\mathbf{y}^{k+1} = \mathbf{w}^k - \frac{1}{L}\mathbf{A}\mathbf{u}^k + \frac{1}{L}\min\{\mathbf{A}\mathbf{u}^k - L\mathbf{w}^k, \mathbf{b}\}$;
  (c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
  (d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k-1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.

Example $1\frac{1}{2}$: Orthogonal Projection onto the Intersection of Closed Convex Sets

$$(P_2) \quad \min_{\mathbf{x}\in\mathbb{E}} \left\{ \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|^2 : \mathbf{x} \in \cap_{i=1}^p C_i \right\}.$$

- $C_1, C_2, \ldots, C_p \subseteq \mathbb{E}$ closed and convex.
- $\mathbf{d} \in \mathbb{E}$.
- Assume that $\cap_{i=1}^p C_i \neq \emptyset$ and that projecting onto each set C_i is an easy task.
- (P_2) fits model (P) with
  $\mathbb{V} = \mathbb{E}^p, f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|^2, g(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) = \sum_{i=1}^p \delta_{C_i}(\mathbf{x}_i)$ and
  $\mathcal{A} : \mathbb{E} \to \mathbb{V}, \mathcal{A}(\mathbf{z}) = (\underbrace{\mathbf{z}, \mathbf{z}, \ldots, \mathbf{z}}_{p \text{ times}})$
- $\underset{\mathbf{x}}{\operatorname{argmax}}\{\langle \mathbf{v}, \mathbf{x}\rangle - f(\mathbf{x})\} = \mathbf{v} + \mathbf{d}$ for any $\mathbf{v} \in \mathbb{E}$;
- $\|\mathcal{A}\|^2 = p$;
- $\sigma = 1$;
- $\mathcal{A}^T(\mathbf{y}) = \sum_{i=1}^p y_i$ for any $\mathbf{y} \in \mathbb{E}^p$;
- $\operatorname{prox}_{Lg}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p) = (P_{C_1}(\mathbf{v}_1), P_{C_2}(\mathbf{v}_2), \ldots, P_{C_p}(\mathbf{v}_p))$ for any $\mathbf{v} \in \mathbb{E}^p$.

DPG and FDPG for Solving (P_2)

Algorithm 3 [DPG for solving (P_2)]

- **Initialization:** $L \geq p, \mathbf{y}^0 \in \mathbb{E}^p$.
- **General Step ($k \geq 0$):**
  (a) $\mathbf{x}^k = \sum_{i=1}^p \mathbf{y}_i^k + \mathbf{d}$;
  (b) $\mathbf{y}_i^{k+1} = \mathbf{y}_i^k - \frac{1}{L}\mathbf{x}^k + \frac{1}{L} P_{C_i}(\mathbf{x}^k - L\mathbf{y}_i^k),\ i = 1, 2, \ldots, p.$

Algorithm 4 [FDPG for solving (P_2)]

- **Initialization:** $L \geq p, \mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{E}^p, t_0 = 1$.
- **General Step ($k \geq 0$):**
  (a) $\mathbf{u}^k = \sum_{i=1}^p \mathbf{w}_i^k + \mathbf{d}$;
  (b) $\mathbf{y}_i^{k+1} = \mathbf{w}_i^k - \frac{1}{L}\mathbf{u}^k + \frac{1}{L} P_{C_i}(\mathbf{u}^k - L\mathbf{w}_i^k)$,
  $i = 1, 2, \ldots, p$;
  (c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
  (d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.

Orthogonal Projection onto a Polyhedral Set Revisited

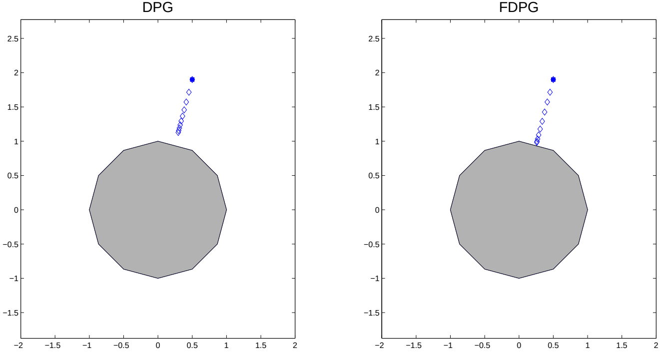
- Algorithm 4 can also be used to find an orthogonal projection of a point $\mathbf{d} \in \mathbb{R}^n$ onto the polyhedral set $C = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$, where $\mathbf{A} \in \mathbb{R}^{p \times n}, \mathbf{b} \in \mathbb{R}^p$.
- C can be written as $C = \cap_{i=1}^p C_i$, where $C_i = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}_i^T \mathbf{x} \leq b_i\}$ with $\mathbf{a}_1^T, \mathbf{a}_2^T, \ldots, \mathbf{a}_p^T$ being the rows of $\mathbf{A}$.
- $P_{C_i}(\mathbf{x}) = \mathbf{x} - \frac{[\mathbf{a}_i^T \mathbf{x} - b_i]_+}{\|\mathbf{a}_i\|^2} \mathbf{a}_i$.

Algorithm 5 [FDPG for solving (P_1)]

- **Initialization:** $L \geq p, \mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{E}^p, t_0 = 1$.
- **General Step ($k \geq 0$):**
  (a) $\mathbf{u}^k = \sum_{i=1}^p \mathbf{w}_i^k + \mathbf{d}$;
  (b) $\mathbf{y}_i^{k+1} = -\frac{1}{L\|\mathbf{a}_i\|^2}[\mathbf{a}_i^T(\mathbf{u}^k - L\mathbf{w}_i^k) - b_i]_+ \mathbf{a}_i, \ i = 1, 2, \ldots, p$;
  (c) $t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$;
  (d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.

Comparison Between DPG and FDPG – Numerical Example

- Consider the problem of projecting the point $(0.5, 1.9)^T$ onto a dodecagon - a regular polygon with 12 edges represented as the intersection of 12 half-spaces.
- The first 10 iterations of the DPG (Algorithm 3) and FDPG (Algorithm 4/5) methods with $L = p = 12$ can be seen below.



Example 2: One-Dimensional Total Variation Denoising

$$(P_3) \quad \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|_2^2 + \lambda\|\mathbf{D}\mathbf{x}\|_1 \right\},$$

- Fits model (P) with
  $\mathbb{E}=\mathbb{R}^n, \mathbb{V}=\mathbb{R}^{n-1}, f(\mathbf{x})=\frac{1}{2}\|\mathbf{x}-\mathbf{d}\|^2 (\sigma=1), g(\mathbf{y})=\lambda\|\mathbf{y}\|_1,\ \mathcal{A}(\mathbf{x})\equiv\mathbf{D}\mathbf{x}$
- $\underset{\mathbf{x}}{\operatorname{argmax}}\{\langle\mathbf{v},\mathbf{x}\rangle - f(\mathbf{x})\} = \mathbf{v}+\mathbf{d}$ for any $\mathbf{v}\in\mathbb{E}$;
- $\|\mathcal{A}\|^2 = \|\mathbf{D}\|_{2,2}^2 \le 4$;
- $\sigma = 1$;
- $\mathcal{A}^T(\mathbf{y}) = \mathbf{D}^T\mathbf{y}$ for any $\mathbf{y}\in\mathbb{R}^{n-1}$;
- $\mathrm{prox}_{Lg}(\mathbf{y}) = \mathcal{T}_{\lambda L}(\mathbf{y})$.

Example 3 Contd.

Algorithm 6 [DPG for solving (P_3)]

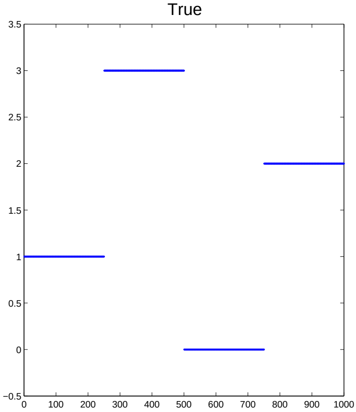
- **Initialization:** $\mathbf{y}^0 \in \mathbb{R}^{n-1}$.
- **General Step ($k \geq 0$):**
  (a) $\mathbf{x}^k = \mathbf{D}^T\mathbf{y}^k + \mathbf{d}$;
  (b) $\mathbf{y}^{k+1} = \mathbf{y}^k - \frac{1}{4}\mathbf{D}\mathbf{x}^k + \frac{1}{4}\mathcal{T}_{4\lambda}(\mathbf{D}\mathbf{x}^k - 4\mathbf{y}^k)$.

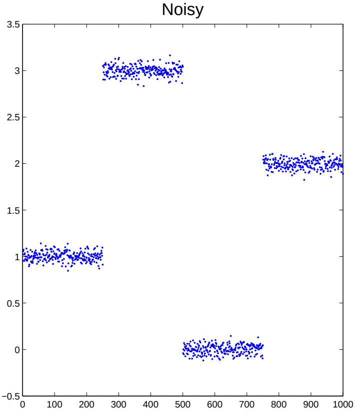
Algorithm 7 [FDPG for solving (P_3)]

- **Initialization:** $\mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{R}^{n-1}$, $t_0 = 1$.
- **General Step ($k \geq 0$):**
  (a) $\mathbf{u}^k = \mathbf{D}^T\mathbf{w}^k + \mathbf{d}$;
  (b) $\mathbf{y}^{k+1} = \mathbf{w}^k - \frac{1}{4}\mathbf{D}\mathbf{u}^k + \frac{1}{4}\mathcal{T}_{4\lambda}(\mathbf{D}\mathbf{u}^k - 4\mathbf{w}^k)$;
  (c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
  (d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k-1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.

Numerical Example

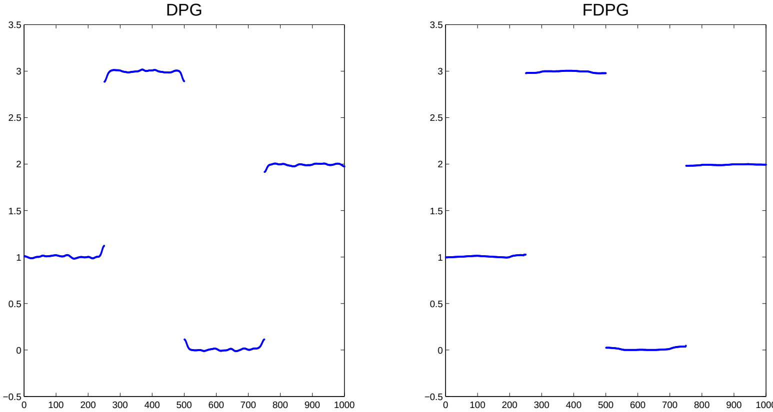
- $n = 1000$
- $\mathbf{d}$ is a noisy measurement of a step function.





Numerical Example Contd.

- 100 iterations of Algorithms 6 (DPG) and 7 (FDPG) initialized with $\mathbf{y}^0 = \mathbf{0}$.



- Objective function values of the DPG and FDPG methods after 100 iterations are 9.1667 and 8.4621 respectively; the optimal value is 8.3031.

The Dual Block Proximal Gradient Method

The Model

$$(Q) \quad \min_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}) + \textstyle\sum_{i=1}^{p} g_i(\mathbf{x}) \right\}.$$

Underlying Assumptions.

(A) $f : \mathbb{E} \to (-\infty, +\infty]$ is proper closed and σ-strongly convex ($\sigma > 0$).

(B) $g_i : \mathbb{E} \to (-\infty, +\infty]$ is proper closed and convex for any $i \in \{1, 2, \ldots, p\}$.

(C) $\mathrm{ri}(\mathrm{dom}(f)) \cap (\cap_{i=1}^{p} \mathrm{ri}(\mathrm{dom}(g_i))) \neq \emptyset$.

Problem (Q) fits model (P) with
$\mathbb{V} = \mathbb{E}^p, g(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) = \sum_{i=1}^{p} g_i(\mathbf{x}_i), \mathcal{A}(\mathbf{z}) = (\underbrace{\mathbf{z}, \mathbf{z}, \ldots, \mathbf{z}}_{p \text{ times}})$.

- $\|\mathcal{A}\|^2 = p$;
- $\mathcal{A}^T(\mathbf{y}) = \sum_{i=1}^{p} y_i$ for any $\mathbf{y} \in \mathbb{E}^p$;
- $\mathrm{prox}_{Lg}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p) = (\mathrm{prox}_{Lg_1}(\mathbf{v}_1), \mathrm{prox}_{Lg_2}(\mathbf{v}_2), \ldots, \mathrm{prox}_{Lg_p}(\mathbf{v}_p))$

FDPG for Solving (Q)

Algorithm 9 [FDPG for solving (Q)]

- **Initialization:** $\mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{E}^p$, $t_0 = 1$.
- **General Step** $(k \geq 0)$**:**

(a) $\mathbf{u}^k = \operatorname{argmax}_{\mathbf{u} \in \mathbb{E}} \left\{ \left\langle \mathbf{u}, \sum_{i=1}^p \mathbf{w}_i^k \right\rangle - f(\mathbf{u}) \right\}$;

(b) $\mathbf{y}_i^{k+1} = \mathbf{w}_i^k - \frac{\sigma}{p}\mathbf{u}^k + \frac{\sigma}{p}\mathrm{prox}_{\frac{p}{\sigma}g_i}(\mathbf{u}^k - \frac{p}{\sigma}\mathbf{w}_i^k)$, $i = 1, 2, \ldots, p$;

(c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;

(d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.

The Dual Block Proximal Gradient Method

- A major disadvantage of Algorithm 9 is the stepsize it uses.
- A way to circumvent this drawback is to employ a dual block proximal gradient method.
- A dual problem to (Q):

$$(DQ) \quad q_{\rm opt} = \max_{\mathbf{y}\in\mathbb{E}^p} \left\{ -f^*\left(\sum_{i=1}^p \mathbf{y}_i\right) - \sum_{i=1}^p \underbrace{g_i^*(-\mathbf{y}_i)}_{G_i(\mathbf{y}_i)} \right\}.$$

- Suppose that the current point is $\mathbf{y}^k = (\mathbf{y}_1^k, \mathbf{y}_2^k, \ldots, \mathbf{y}_p^k)$. At each iteration we pick an index i according to some rule and perform a proximal gradient step on ith block:

$$\mathbf{y}_i^{k+1} = \mathrm{prox}_{\sigma G_i}\left(\mathbf{y}_i^k - \sigma \nabla f^*\left(\sum_{j=1}^p \mathbf{y}_j^k\right)\right).$$

Dual Representation

The Dual Block Proximal Gradient (DBPG) Method – dual representation

- **Initialization:** pick $\mathbf{y}^0 = (\mathbf{y}_1^0, \mathbf{y}_2^0, \ldots, \mathbf{y}_p^0) \in \mathbb{E}^p$.
- **General step** $(k \geq 0)$**:**
  - pick an index $i_k \in \{1, 2, \ldots, p\}$;
  - compute $\mathbf{y}_j^{k+1} = \begin{cases} \mathrm{prox}_{\sigma G_{i_k}}\left(\mathbf{y}_{i_k}^k - \sigma \nabla f^*(\sum_{j=1}^p \mathbf{y}_j^k)\right), & j = i_k, \\ \mathbf{y}_j^k, & j \neq i_k. \end{cases}$

Lemma. The relation $\mathbf{y}_i = \mathrm{prox}_{\frac{1}{L} G_i}\left(\mathbf{v}_i - \frac{1}{L}\nabla f^*(\sum_{j=1}^p \mathbf{v}_j)\right)$ holds if and only if

$$\mathbf{y}_i = \mathbf{v}_i - \frac{1}{L}\tilde{\mathbf{x}} + \frac{1}{L}\mathrm{prox}_{Lg_i}\left(\tilde{\mathbf{x}} - L\mathbf{v}_i\right),$$

where $\tilde{\mathbf{x}} = \underset{\mathbf{x} \in \mathbb{E}}{\operatorname{argmax}} \left\{ \langle \mathbf{x}, \sum_{j=1}^p \mathbf{v}_j \rangle - f(\mathbf{x}) \right\}$.

Primal Representation

The Dual Block Proximal Gradient (DBPG) Method – primal representation

Initialization. pick $\mathbf{y}^0 = (\mathbf{y}_1^0, \mathbf{y}_2^0, \ldots, \mathbf{y}_p^0) \in \mathbb{E}$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick $i_k \in \{1, 2, \ldots, p\}$.

(b) set $\mathbf{x}^k = \underset{\mathbf{x}\in\mathbb{E}}{\operatorname{argmax}} \left\{ \langle \mathbf{x}, \sum_{j=1}^p \mathbf{y}_j^k \rangle - f(\mathbf{x}) \right\}$.

(c) set $\mathbf{y}_j^{k+1} = \begin{cases} \mathbf{y}_{i_k}^k - \sigma\mathbf{x}^k + \sigma\mathrm{prox}_{g_i/\sigma}\left(\mathbf{x}^k - \mathbf{y}_{i_k}^k/\sigma\right), & j = i_k, \\ \mathbf{y}_j^k, & j \neq i_k. \end{cases}$

Possible stepsize strategies.

- **cyclic.** $i_k = (k \mod p) + 1$.
- **randomized.** i_k is randomly picked from $\{1, 2, \ldots, p\}$ by a uniform distribution.

Rates of Convergence of the Cyclic and Randomized DBPG Methods

- $O(1/k)$ rates of convergence of the sequences of dual objective function values follow by the corresponding results on the block proximal gradient method.
- $O(1/k)$ rates of the primal sequence follow by the primal-dual relation.

Cyclic:

(a) $q_{\rm opt} - q(\mathbf{y}^{pk}) \le \max\left\{\left(\frac{1}{2}\right)^{(k-1)/2}(q_{\rm opt} - q(\mathbf{y}^0)), \frac{8p(p+1)^2R^2}{\sigma(k-1)}\right\}.$

(b) $\|\mathbf{x}^{pk} - \mathbf{x}^*\|^2 \le \frac{2}{\sigma}\max\left\{\left(\frac{1}{2}\right)^{(k-1)/2}(q_{\rm opt} - q(\mathbf{y}^0)), \frac{8p(p+1)^2R^2}{\sigma(k-1)}\right\}.$

Randomized:

(a) $q_{\rm opt} - \mathsf{E}_{\xi_k}(q(\mathbf{y}^{k+1})) \le \frac{p}{p+k+1}\left(\frac{1}{2\sigma}\|\mathbf{y}^0 - \mathbf{y}^*\|^2 + q_{\rm opt} - q(\mathbf{y}^0)\right).$

(b) $\mathsf{E}_{\xi_k}\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 \le \frac{2p}{\sigma(p+k+1)}\left(\frac{1}{2\sigma}\|\mathbf{y}^0 - \mathbf{y}^*\|^2 + q_{\rm opt} - q(\mathbf{y}^0)\right).$

THE END

THANK YOU FOR YOUR ATTENTION