10 Lectures on "Proximal-Based Methods"

Amir Beck

Mini-Course Given at the Vienna School On Computational Optimization
17-21 October, 2016

Course Overview

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

- ► 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 Course 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

Lecture 1 - Vector Spaces

a vector space $\mathbb{E}$ over $\mathbb{R}$ (or a "real vector space") is a set of elements called vectors such that the following holds.

(A) for any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{E}$, there corresponds a vector $\mathbf{x} + \mathbf{y}$, called the sum of $\mathbf{x}$ and $\mathbf{y}$ satisfying the following properties:

    1. $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$.
    2. $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$ for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{E}$.
    3. $\exists \mathbf{0} \in \mathbb{E}$ (zeros vector) such that $\mathbf{x} + \mathbf{0} = \mathbf{x}$ for any $\mathbf{x}$.
    4. $\forall \mathbf{x} \in \mathbb{E}$ there exists a vector $-\mathbf{x} \in \mathbb{E}$ such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$.

(B) for any real number (also called scalar) $\alpha \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{E}$, there corresponds a vector $\alpha \mathbf{x}$ called the scalar multiplication of α and $\mathbf{x}$ satisfying the following properties:

    1. $\alpha(\beta \mathbf{x}) = (\alpha \beta)\mathbf{x}$ for any $\alpha, \beta \in \mathbb{R}, \mathbf{x} \in \mathbb{E}$.
    2. $1\mathbf{x} = \mathbf{x}$ for any $\mathbf{x} \in \mathbb{E}$.

(C) the two operations (summation, scalar multiplication) satisfy

    1. $\alpha(\mathbf{x} + \mathbf{y}) = \alpha \mathbf{x} + \alpha \mathbf{y}$ for any $\alpha \in \mathbb{R}, \mathbf{x}, \mathbf{y} \in \mathbb{E}$.
    2. $(\alpha + \beta)\mathbf{x} = \alpha \mathbf{x} + \beta \mathbf{y}$.

Dimension

▶ a set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\} \subseteq \mathbb{E}$ is linearly independent if

$$\sum_{i=1}^{n} \alpha_i \mathbf{v}_i = \mathbf{0} \Rightarrow \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$$

▶ $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is said to span $\mathbb{E}$ if for any $\mathbf{x} \in \mathbb{E}$, there exists $\beta_1, \beta_2, \ldots, \beta_n \in \mathbb{R}$ such that

$$\mathbf{x} = \sum_{i=1}^{n} \beta_i \mathbf{v}_i.$$

▶ a basis of $\mathbb{E}$ is an independent set of vectors that spans $\mathbb{E}$.

▶ the number of vectors in all the bases of a vector space $\mathbb{E}$ is the same and called the dimension of $\mathbb{E}$. Notation: $\dim(\mathbb{E})$.

▶ we will only consider finite-dimensional vector spaces.

Norms

- a norm $\|\cdot\|$ on a vector space $\mathbb{E}$ is a function $\|\cdot\| : \mathbb{E} \to \mathbb{R}$ satisfying
  1. **(nonnegativity)** $\|\mathbf{x}\| \geq 0$ for any $\mathbf{x} \in \mathbb{E}$ and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
  2. **(positive homogeneity)** $\|\lambda\mathbf{x}\| = |\lambda| \cdot \|\mathbf{x}\|$ for any $\mathbf{x} \in \mathbb{E}$ and $\lambda \in \mathbb{R}$.
  3. **(triangle inequality)** $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$.

- **open balls:** $(\mathbf{c} \in \mathbb{E}, r > 0)$

$$B(\mathbf{c}, r) = \{\mathbf{x} \in \mathbb{E} : \|\mathbf{x} - \mathbf{c}\| < r\}.$$

- **closed balls:** $(\mathbf{c} \in \mathbb{E}, r > 0)$

$$B[\mathbf{c}, r] = \{\mathbf{x} \in \mathbb{E} : \|\mathbf{x} - \mathbf{c}\| \leq r\}.$$

- we will sometimes use the notation $B_{\|\cdot\|}[\mathbf{c}, r]$ or $B_{\|\cdot\|}(\mathbf{c}, r)$ to identify the specific norm that is being used.

Inner Products

▶ an inner product of a real vector space $\mathbb{E}$ is a function that associates to each pair of vectors $\mathbf{x}, \mathbf{y}$ a real number, which is denoted by $\langle \mathbf{x}, \mathbf{y} \rangle$ and satisfies the following properties:

1. **(commutativity)** $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$.
2. **(linearity)** $\langle \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2, \mathbf{y} \rangle = \alpha_1 \langle \mathbf{x}_1, \mathbf{y} \rangle + \alpha_2 \langle \mathbf{x}_2, \mathbf{y} \rangle$ for any $\alpha_1, \alpha_2 \in \mathbb{R}$ and $\mathbf{x}_1, \mathbf{x}_2, \mathbf{y} \in \mathbb{E}$.
3. **(positive definiteness)** $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ for any $\mathbf{x} \in \mathbb{E}$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

▶ a vector space endowed with an inner product is also called an inner product space.

we will always 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 $\| \cdot \|$.

Affine Sets

- $S \subseteq \mathbb{E}$ is called affine if $\mathbf{x}, \mathbf{y} \in S, \lambda \in \mathbb{R} \Rightarrow \lambda \mathbf{x} + (1 - \lambda)\mathbf{y} \in S$.
- for a set $S \subseteq \mathbb{E}$, the affine hull of S, denoted by $\text{aff}(S)$, is the intersection of all affine sets containing S.
- $\text{aff}(S)$ is an affine set, and is the smallest affine set containing S (w.r.t. inclusion).
- hyperplanes: $(\mathbf{0} \neq \mathbf{a} \in \mathbb{E}, b \in \mathbb{R})$

$$H_{\mathbf{a},b} = \{\mathbf{x} \in \mathbb{E} : \langle \mathbf{a}, \mathbf{x} \rangle = b\},$$

hyperplanes are affine sets.

Convex Sets

- a set $C \subseteq \mathbb{E}$ is called convex if
  $\mathbf{x}, \mathbf{y} \in C, \lambda \in [0,1] \Rightarrow \lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in C$.
- affine sets are convex.
- open and closed balls are convex.
- **closed line segments** $(\mathbf{x}, \mathbf{y} \in \mathbb{E})$

$$[\mathbf{x}, \mathbf{y}] = \{\alpha\mathbf{x} + (1-\alpha)\mathbf{y} : \alpha \in [0,1]\}.$$

- **open line segments** $(\mathbf{x}, \mathbf{y} \in \mathbb{E})$

$$(\mathbf{x}, \mathbf{y}) = \{\alpha\mathbf{x} + (1-\alpha)\mathbf{y} : \alpha \in (0,1)\}.$$

- **half spaces:** $(\mathbf{a} \in \mathbb{E}, b \in \mathbb{R})$

$$H_{\mathbf{a},b}^- = \{\mathbf{x} \in \mathbb{E} : \langle \mathbf{a}, \mathbf{x} \rangle \le b\}.$$

half spaces are convex.

Euclidean Spaces

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

The Space $\mathbb{R}^n$

▶ the vector space $\mathbb{R}^n$ is the set of n-dimensional column vectors with real components endowed with the component-wise addition operator:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix},$$

and the scalar-vector product

$$\lambda \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \lambda x_1 \\ \lambda x_2 \\ \vdots \\ \lambda x_n \end{pmatrix}.$$

▶ the standard basis of $\mathbb{R}^n$ is denoted by $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$.
▶ $\mathbf{e}$ - vector of all ones, $\mathbf{0}$ - vector of all zeros.

The Space $\mathbb{R}^n$

- most popular inner product – the dot product:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i.$$

> unless otherwise stated, we will assume that the endowed inner product in $\mathbb{R}^n$ is the dot product.

- **Q-inner product**: $(\mathbf{Q} \succ \mathbf{0})$

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

- **I**-inner product is the dot product.
- Euclidean norm associated with the dot product.

$$\textit{l}_2\textbf{-norm: } \|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

- Euclidean norm associated with the **Q**-inner product.

$$\textbf{Q-norm: } \|\mathbf{x}\|_{\mathbf{Q}} = \sqrt{\mathbf{x}^T \mathbf{Q} \mathbf{x}}.$$

The Space $\mathbb{R}^n$

- l_p-**norm:** $(p \geq 1)$

$$\|\mathbf{x}\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}.$$

- l_∞-**norm**:

$$\|\mathbf{x}\|_\infty = \max_{i=1,2,\ldots,n} |x_i|.$$

Subsets of $\mathbb{R}^n$

▶ **nonnegative orthant:**
$$\mathbb{R}^n_+ = \left\{ (x_1, x_2, \ldots, x_n)^T : x_1, x_2, \ldots, x_n \geq 0 \right\}.$$

▶ **positive orthant:**
$$\mathbb{R}^n_{++} = \left\{ (x_1, x_2, \ldots, x_n)^T : x_1, x_2, \ldots, x_n > 0 \right\}.$$

▶ **unit-simplex:**
$$\Delta_n = \left\{ \mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq \mathbf{0}, \mathbf{e}^T \mathbf{x} = 1 \right\}.$$

▶ **box:** $(\boldsymbol{\ell} \leq \mathbf{u})$
$$\mathrm{Box}[\boldsymbol{\ell}, \mathbf{u}] = \{ \mathbf{x} \in \mathbb{R}^n : \boldsymbol{\ell} \leq \mathbf{x} \leq \mathbf{u} \}.$$

Operations on Vectors in $\mathbb{R}^n$

$\mathbf{a}, \mathbf{b}, \mathbf{x} \in \mathbb{R}^n$.

- **nonnegative part:** $[\mathbf{x}]_+ = (\max\{x_i, 0\})_{i=1}^n$
- **absolute value:** $|\mathbf{x}| = (|x_i|)_{i=1}^n$
- **sign:**

$$\mathrm{sgn}(\mathbf{x})_i = \begin{cases} 1, & x_i \geq 0, \\ -1, & x_i < 0. \end{cases}$$

- **Hadmard product:** $\mathbf{a} \odot \mathbf{b} = (a_i b_i)_{i=1}^n$

The space $\mathbb{R}^{m \times n}$

- the set of all real-valued $m \times n$ matrices is denote by $\mathbb{R}^{m \times n}$.
- $\mathbb{R}^{m \times n}$ a vector space with the component-wise addition as the summation operation and the component-wise scalar multiplication as the "scalar-vector multiplication" operation.
- the dot product in $\mathbb{R}^{m \times n}$ is defined by

$$\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{A}^T \mathbf{B}) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij}, \quad \mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}.$$

unless otherwise stated, the endowed inner product in $\mathbb{R}^{m \times n}$ is the dot product.

Subsets of $\mathbb{R}^{n \times n}$

- **symmetric matrices:**

$$\mathbb{S}^n = \left\{ \mathbf{A} \in \mathbb{R}^{n \times n} : \mathbf{A} = \mathbf{A}^T \right\}.$$

$\mathbb{S}^n$ is also a vector space.

- **positive semidefinite matrices:**

$$\mathbb{S}^n_+ = \left\{ \mathbf{A} \in \mathbb{R}^{n \times n} : \mathbf{A} \succeq \mathbf{0} \right\},$$

- **positive definite matrices:**

$$\mathbb{S}^n_{++} = \left\{ \mathbf{A} \in \mathbb{R}^{n \times n} : \mathbf{A} \succ \mathbf{0} \right\}.$$

- **orthogonal matrices:**

$$\mathbb{O}_n = \left\{ \mathbf{A} \in \mathbb{R}^{n \times n} : \mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I} \right\}.$$

Norms in $\mathbb{R}^{m \times n}$

- if $\mathbb{R}^{m \times n}$ is endowed with the inner product, then the corresponding Euclidean norm is the Frobenius norm:

$$\|\mathbf{A}\|_F = \mathrm{Tr}(\mathbf{A}^T \mathbf{A}) = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2}, \quad \mathbf{A} \in \mathbb{R}^{m \times n}.$$

- induced norms: given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, the induced matrix norm is

$$\|\mathbf{A}\|_{a,b} = \max_{\mathbf{x}} \{\|\mathbf{A}\mathbf{x}\|_b : \|\mathbf{x}\|_a \leq 1\}$$

- trivial inequality: $\|\mathbf{A}\mathbf{x}\|_b \leq \|\mathbf{A}\|_{a,b} \|\mathbf{x}\|_a$
- **spectral norm:** $\|\mathbf{A}\|_{2,2} = \sqrt{\lambda_{\max}(\mathbf{A}^T \mathbf{A})} \equiv \sigma_{\max}(\mathbf{A})$.
- **maximum absolute column sum norm:**
  $\|\mathbf{A}\|_{1,1} = \max_{j=1,2,\ldots,n} \sum_{i=1}^{m} |A_{i,j}|$.
- **maximum absolute row sum norm:**
  $\|\mathbf{A}\|_{\infty,\infty} = \max_{i=1,2,\ldots,m} \sum_{j=1}^{n} |A_{i,j}|$.

The Dual Norm

▶ Given a norm $\|\cdot\|$, which is not necessarily the endowed Euclidean norm, the dual norm is defined as

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

▶ Alternative representation:

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

Lemma. **(Generalized Cauchy-Schwarz Inequality)** Let $\mathbb{E}$ be an inner product vector space, and let $\|\cdot\|$ be a norm. Then

$$|\langle \mathbf{y}, \mathbf{x}\rangle| \leq \|\mathbf{y}\|_*\|\mathbf{x}\| \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{E}$$

Examples: Euclidean norms are self-dual; in $\mathbb{R}^n$, the dual of the l_p-norm ($p \in [1, \infty]$) is the l_q norm ($\frac{1}{p} + \frac{1}{q} = 1$); the dual of the **Q**-norm is the **Q**$^{-1}$-norm.

Cartesian Product of Vector Spaces

given m vector spaces $\mathbb{E}_1, \mathbb{E}_2, \ldots, \mathbb{E}_m$, equipped with inner products $\langle \cdot, \cdot \rangle_{\mathbb{E}_i}$, their Cartesian product $\mathbb{E}_1 \times \mathbb{E}_2 \cdots \times \mathbb{E}_m$ is the vector space of all m-tuples $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m), \mathbf{v}_i \in \mathbb{E}_i$ equipped with

▶ componentwise addition between vectors:

$$(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m) + (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m) = (\mathbf{v}_1 + \mathbf{w}_1, \mathbf{v}_2 + \mathbf{w}_2, \ldots, \mathbf{v}_m + \mathbf{w}_m)$$

▶ scalar-vector multiplication operation:

$$\alpha(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m) = (\alpha\mathbf{v}_1, \alpha\mathbf{v}_2, \ldots, \alpha\mathbf{v}_m).$$

▶ inner product:

$$\langle (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m), (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m) \rangle_{\mathbb{E}_1 \times \mathbb{E}_2 \times \cdots \mathbb{E}_m} = \sum_{i=1}^{m} \langle \mathbf{v}_i, \mathbf{w}_i \rangle_{\mathbb{E}_i}.$$

Norms on Cartesian Product of Vector Spaces

suppose that $\mathbb{E}_1, \mathbb{E}_2, \ldots, \mathbb{E}_m$ are vector spaces with endowed norms $\|\cdot\|_{\mathbb{E}_1}, \|\cdot\|_{\mathbb{E}_2}, \ldots, \|\cdot\|_{\mathbb{E}_m}$ respectively. Possible norms on $\mathbb{E}_1 \times \mathbb{E}_2 \times \cdots \times \mathbb{E}_m$:

▶ **composite l_p norm:** $(p \geq 1)$

$$\|(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m)\| = \sqrt[p]{\sum_{i=1}^m \|\mathbf{u}_i\|_{\mathbb{E}_i}^p}$$

▶ **composite weighted l_2 norm:** $(\omega_i > 0)$

$$\|(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m)\| = \sqrt{\sum_{i=1}^m \omega_i \|\mathbf{u}_i\|_{\mathbb{E}_i}^2}$$

▶ If $\mathbb{E}_1, \mathbb{E}_2, \ldots, \mathbb{E}_m$ are Euclidean spaces, we will use the convention that then $\mathbb{E}_1 \times \mathbb{E}_2 \times \cdots \times \mathbb{E}_m$ is also an Euclidean space and

$$\|(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m)\|_{\mathbb{E}_1 \times \mathbb{E}_2 \times \cdots \mathbb{E}_m} = \sqrt{\sum_{i=1}^m \|\mathbf{u}_i\|_{\mathbb{E}_i}^2}$$

Linear Transformations

▶ given two vector spaces $\mathbb{E}$ and $\mathbb{V}$, a function $\mathcal{A} : \mathbb{E} \to \mathbb{V}$ is called a linear transformation if

$$\mathcal{A}(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\mathcal{A}(\mathbf{x}) + \beta\mathcal{A}(\mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{E}, \alpha, \beta \in \mathbb{R}$$

▶ all linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$ have the form

$$\mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x},$$

for some matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$.

▶ All linear transformations from $\mathbb{R}^{m \times n}$ to $\mathbb{R}^k$ have the form

$$\mathcal{A}(\mathbf{X}) = \begin{pmatrix} \mathrm{Tr}(\mathbf{A}_1^T \mathbf{X}) \\ \mathrm{Tr}(\mathbf{A}_2^T \mathbf{X}) \\ \vdots \\ \mathrm{Tr}(\mathbf{A}_k^T \mathbf{X}) \end{pmatrix}$$

for some $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k \in \mathbb{R}^{m \times n}$.

▶ $\mathcal{I}$ is the identity transformation defined by the relation $\mathcal{I}(\mathbf{x}) = \mathbf{x}$ for all $\mathbf{x} \in \mathbb{E}$.

Adjoint Transformations

▶ given a linear transformation $\mathcal{A} : \mathbb{V} \to \mathbb{E}$, where $\mathbb{E}$ and $\mathbb{V}$ are Euclidean, the adjoint transformation denoted by $\mathcal{A}^T$ is a transformation from $\mathbb{E}$ to $\mathbb{V}$ defined by the relation

$$\langle \mathbf{y}, \mathcal{A}(\mathbf{x}) \rangle = \langle \mathcal{A}^T(\mathbf{y}), \mathbf{x} \rangle$$

for any $\mathbf{x} \in \mathbb{V}, \mathbf{y} \in \mathbb{E}$.

▶ when $\mathbb{V} = \mathbb{R}^n, \mathbb{E} = \mathbb{R}^m$, and $\mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ for $\mathbf{A} \in \mathbb{R}^{m \times n}$, then $\mathcal{A}^T(\mathbf{x}) = \mathbf{A}^T\mathbf{x}$.

▶ Let $\mathcal{A} : \mathbb{R}^{m \times n} \to \mathbb{R}^k$ be given by

$$\mathcal{A}(\mathbf{X}) = \begin{pmatrix} \mathrm{Tr}(\mathbf{A}_1^T\mathbf{X}) \\ \mathrm{Tr}(\mathbf{A}_2^T\mathbf{X}) \\ \vdots \\ \mathrm{Tr}(\mathbf{A}_k^T\mathbf{X}) \end{pmatrix},$$

where $\mathbf{A}_i \in \mathbb{R}^{m \times n}$ are given matrices. Then $\mathcal{A}^T(\mathbf{y}) = \sum_{i=1}^k y_i \mathbf{A}_i$.

▶ **useful properties:** $(\mathcal{A}^T)^T = \mathcal{A}$ and $(\mathcal{A}^T)^{-1} = (\mathcal{A}^{-1})^T$ (if $\mathcal{A}$ inv.)

Norms of Linear Transformation

- Let $\mathcal{A} : \mathbb{E} \to \mathbb{V}$. The norm of the linear transformation is defined by

$$\|\mathcal{A}\| \equiv \max\{\|\mathcal{A}(\mathbf{x})\|_{\mathbb{V}} : \|\mathbf{x}\|_{\mathbb{E}} = 1\}.$$

- **Property:** It is not difficult to show that $\|\mathcal{A}\| = \|\mathcal{A}^T\|$.
- Generalization of the notion of induced norms for $m \times n$ matrices.

References

► P. R. Halmos, *Finite-dimensional vector spaces*, second edition (1974).

Lecture 2–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:**

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

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

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

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

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

Closedness

- The epigraph of an extended real-valued function $f : \mathbb{E} \to [-\infty, \infty]$ is defined by
  $$\mathrm{epi}(f) = \{(\mathbf{x}, y) : f(\mathbf{x}) \leq y, \mathbf{x} \in \mathbb{E}, y \in \mathbb{R}\} \subseteq \mathbb{E} \times \mathbb{R}.$$

- A function $f : \mathbb{R}^n \to [-\infty, \infty]$ 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 $\mathrm{dom}(f) \neq \emptyset$.

- 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.
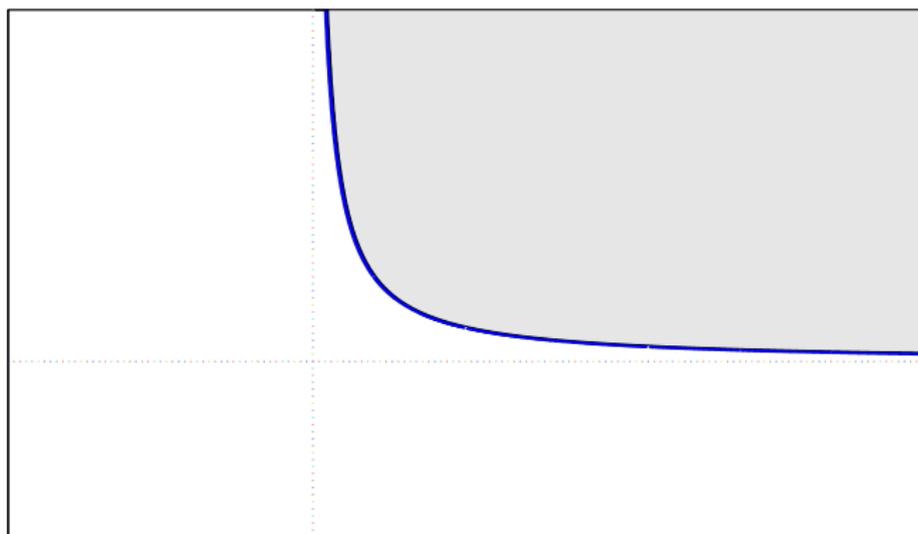
Proof.
$$\mathrm{epi}(f) = \{(\mathbf{x}, y) \in \mathbb{E} \times \mathbb{R} : \delta_C(\mathbf{x}) \leq y\} = C \times \mathbb{R}_+,$$

which is evidently closed if and only if C is closed. $\square$

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. Let $f : \mathbb{E} \to [-\infty, \infty]$. Then the following three 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.

Proof

- **(i $\Rightarrow$ ii)** Suppose that f is l.s.c.. We will show that $\mathrm{epi}(f)$ is closed.
- Take $\{(\mathbf{x}_n, y_n)\}_{n \geq 1} \subseteq \mathrm{epi}(f)$ such that $(\mathbf{x}_n, y_n) \to (\mathbf{x}^*, y^*)$ as $n \to \infty$.
- $f(\mathbf{x}_n) \leq y_n$ for all $n \geq 1$
- $f(\mathbf{x}^*) \leq \liminf_{n \to \infty} f(\mathbf{x}_n) \leq \liminf_{n \to \infty} y_n = y^*$
- Showing that $(\mathbf{x}^*, y^*) \in \mathrm{epi}(f)$, and hence that f is closed.
- **(ii $\Rightarrow$ iii)** Suppose that f is closed. We will show that $\mathrm{Lev}(f, \alpha)$ is closed for all $\alpha \in \mathbb{R}$.
- If $\mathrm{Lev}(f, \alpha) = \emptyset$, we are done. Otherwise, take a sequence $\{\mathbf{x}_n\}_{n \geq 1} \subseteq \mathrm{Lev}(f, \alpha)$ that converges to $\bar{\mathbf{x}}$.
- $(\mathbf{x}_n, \alpha) \in \mathrm{epi}(f)$ for any n and $(\mathbf{x}_n, \alpha) \to (\bar{\mathbf{x}}, \alpha)$ as $n \to \infty$.
- Thus, $(\bar{\mathbf{x}}, \alpha) \in \mathrm{epi}(f) \Rightarrow \bar{\mathbf{x}} \in \mathrm{Lev}(f, \alpha)$.

Proof Contd.

- ▶ **(iii ⇒ i)** Assume all level sets are closed. We will show that f is l.s.c.
- ▶ Assume in contradiction that f is not l.s.c., meaning that $\exists \mathbf{x}^* \in \mathbb{E}$ and $\{\mathbf{x}_n\}_{n \geq 1} \subseteq \mathbb{E}$ such that $\mathbf{x}_n \to \mathbf{x}^*$ and $\liminf_{n \to \infty} f(\mathbf{x}_n) < f(\mathbf{x}^*)$.
- ▶ Take α for which

$$\liminf_{n \to \infty} f(\mathbf{x}_n) < \alpha < f(\mathbf{x}^*). \tag{1}$$

- ▶ $\exists \{\mathbf{x}_{n_k}\}_{k \geq 1}$ converging to $\mathbf{x}^*$ such that $f(\mathbf{x}_{n_k}) \leq \alpha$ for all $k \geq 1$.
- ▶ Then $f(\mathbf{x}^*) \leq \alpha$, which is a contradiction to (1).

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.

Closedness Vs. Continuity

> Theorem Let $f : \mathbb{E} \to [-\infty, \infty]$ be an extended real-valued function that is continuous over its domain, and suppose that $\mathrm{dom}(f)$ is closed. Then f is closed.

Proof. Exercise

Examples

- $f : \mathbb{R} \to (-\infty, \infty]$ is given by

$$
f_\alpha(x) = \left\{ \begin{array}{ll} \alpha, & x = 0, \\ x, & 0 < x \leq 1, \\ \infty, & \text{else}. \end{array} \right.
$$

  for which values of α is the function closed? continuous over its domain?

- Consider 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.

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 $C \cap \text{dom}(f) \neq \emptyset$. Then

(a) f is bounded below over C.

(b) f attains a minimizer over C.

Proof. Exercise

Coerciveness

- 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.

Proof.

- Let $\mathbf{x}_0 \in S \cap \mathrm{dom}(f)$.
- There exists an $M > 0$ for which $f(\mathbf{x}) > f(\mathbf{x}_0)$ for any $\mathbf{x}$ satisfying $\|\mathbf{x}\| > M$
- The set of minimizers of f over S = the set of minimizers of f over $S \cap B_{\|\cdot\|}[\mathbf{0}, M]$, which is compact and nonempty.
- By Weierstrass theorem for closed functions, there exists a global minimizer of f over $S \cap B_{\|\cdot\|}[\mathbf{0}, M]$, and hence also 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.

Given a nonempty set $C \subseteq \mathbb{E}$, the distance function to C is defined by

$$d_C(\mathbf{x}) = \min_{\mathbf{y} \in C} \|\mathbf{x} - \mathbf{y}\|.$$

Example: $\varphi_C(\mathbf{x}) = \frac{1}{2}(\|\mathbf{x}\|^2 - d_C^2(\mathbf{x}))$

Are Closed and Convex Functions Continuous?

Not in general, but it is correct in the 1D case:

> Theorem (continuity of 1D closed convex functions) Let $f : \mathbb{R} \to (-\infty, \infty]$ be a proper closed and convex function whose domain is $\text{dom} f = [a, b]$, where $a, b \in \mathbb{R}, a < b$. Then f is continuous over $[a, b]$.

Proof. technical and long.

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

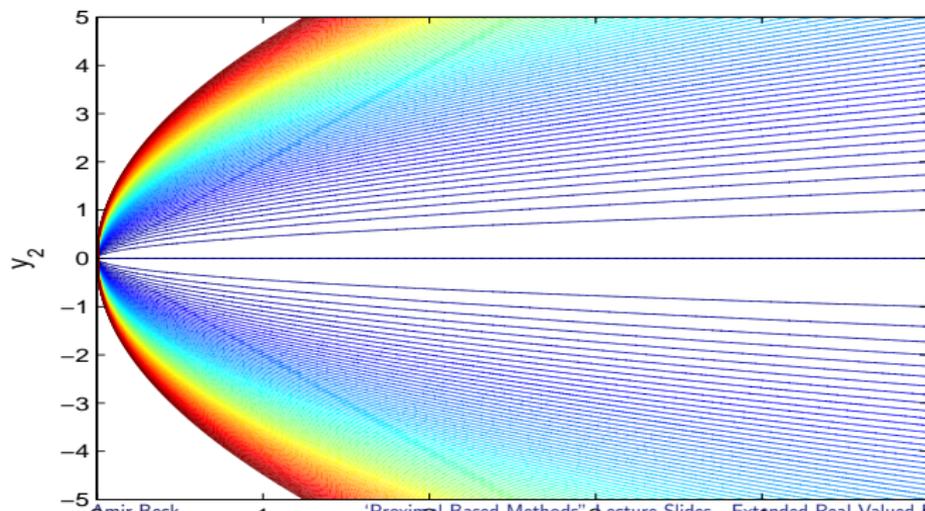| C	$\sigma_C(\mathbf{y})$	assumptions	Example No.
$\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$	$\max_{i=1,2,\ldots,n}\langle\mathbf{b}_i, \mathbf{y}\rangle$	$\mathbf{b}_i \in \mathbb{E}$	1
K	$\delta_{K^\circ}(\mathbf{y})$	K – cone	2
$\mathbb{R}^n_+$	$\delta_{\mathbb{R}^n_-}(\mathbf{y})$	$\mathbb{E}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

A Discontinuous Closed and Convex Function

If

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

Then

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

Support of Closure and Convex Hull

Lemma. Let $A \subseteq \mathbb{E}$ be nonempty. Then

(a) $\sigma_A = \sigma_{\mathrm{cl}(A)}$.

(b) $\sigma_A = \sigma_{\mathrm{conv}(A)}$.

Proof. Exercise

References

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

Lecture 3 - 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 $\partial f(\mathbf{x}) = \emptyset$.

Example 1

▶ Consider the function $f : \mathbb{E} \to \mathbb{R}$ given by $f(\mathbf{x}) = \|\mathbf{x}\|$ ($\|\cdot\|$ being an arbitrary norm).

▶ We will show that

$$\partial f(\mathbf{0}) = B_{\|\cdot\|_*}[\mathbf{0}, 1] = \{\mathbf{y} \in \mathbb{E} : \|\mathbf{y}\|_* \leq 1\}.$$

▶ To show the above, note that $\mathbf{g} \in \partial f(\mathbf{0})$ if and only if

$$f(\mathbf{y}) \geq f(\mathbf{0}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{0} \rangle \text{ for all } \mathbf{y} \in \mathbb{E},$$

which is the same as

$$\|\mathbf{y}\| \geq \langle \mathbf{g}, \mathbf{y} \rangle \text{ for all } \mathbf{y} \in \mathbb{E}. \tag{1}$$

▶ (1) is the same as the statement $\|\mathbf{g}\|_* \leq 1$. Why?

Consequence: Let $h : \mathbb{R}^n \to \mathbb{R}$ be given by $h(\mathbf{x}) = \|\mathbf{x}\|_1$. Then $\partial h(0) = [-1, 1]^n$.

Example 2 - Subdifferential of the Indicator

Let $S \subseteq \mathbb{E}$ be nonempty. We will show that

$$\partial \delta_S(\mathbf{x}) = N_S(\mathbf{x}) \text{ for all } \mathbf{x} \in S$$

where

$N_S(\mathbf{x})$ is the normal cone defined by

$$N_S(\mathbf{x}) = \{\mathbf{y} \in \mathbb{E} : \langle \mathbf{y}, \mathbf{z} - \mathbf{x} \rangle \leq 0 \text{ for any } \mathbf{z} \in S\}.$$

Proof.

- $\mathbf{y} \in \partial \delta_S(\mathbf{x})$ if and only if

$$\delta_S(\mathbf{z}) \geq \delta_S(\mathbf{x}) + \langle \mathbf{y}, \mathbf{z} - \mathbf{x} \rangle \text{ for all } \mathbf{z} \in S,$$

- which is the same as

$$\langle \mathbf{y}, \mathbf{z} - \mathbf{x} \rangle \leq 0, \text{ for all } \mathbf{z} \in S.$$

- that is, $\mathbf{y} \in N_S(\mathbf{x})$. $\square$

For $\mathbf{x} \notin S$, $\partial \delta_S(\mathbf{x}) = \emptyset$ and $N_S(\mathbf{x}) = \emptyset$

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. $\square$

Nonemptiness of Subdifferential Sets $\Rightarrow$ Convexity

> **Lemma.** Let $f : \mathbb{E} \to (-\infty, \infty]$ be an extended real-valued function and assume that $\text{dom}(f)$ is convex. Suppose that for any $\mathbf{x} \in \text{dom}(f)$, the set $\partial f(\mathbf{x})$ is nonempty. Then f is convex.

Proof.

- Let $\mathbf{x}, \mathbf{y} \in \text{dom}(f), \alpha \in [0, 1]$.
- $\mathbf{z}_\alpha = (1 - \alpha)\mathbf{x} + \alpha\mathbf{y} \in \text{dom}(f)$.
- there exists $\mathbf{g} \in \partial f(\mathbf{z}_\alpha)$
-

$$
\begin{aligned}
f(\mathbf{y}) &\geq f(\mathbf{z}_\alpha) + \langle \mathbf{g}, \mathbf{y} - \mathbf{z}_\alpha \rangle = f(\mathbf{z}_\alpha) + (1 - \alpha)\langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle, \\
f(\mathbf{x}) &\geq f(\mathbf{z}_\alpha) + \langle \mathbf{g}, \mathbf{x} - \mathbf{z}_\alpha \rangle = f(\mathbf{z}_\alpha) - \alpha\langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle.
\end{aligned}
$$

- Thus, $f((1 - \alpha)\mathbf{x} + \alpha\mathbf{y}) = f(\mathbf{z}_\alpha) \leq (1 - \alpha)f(\mathbf{x}) + \alpha f(\mathbf{y})$,
- Hence, f is convex.
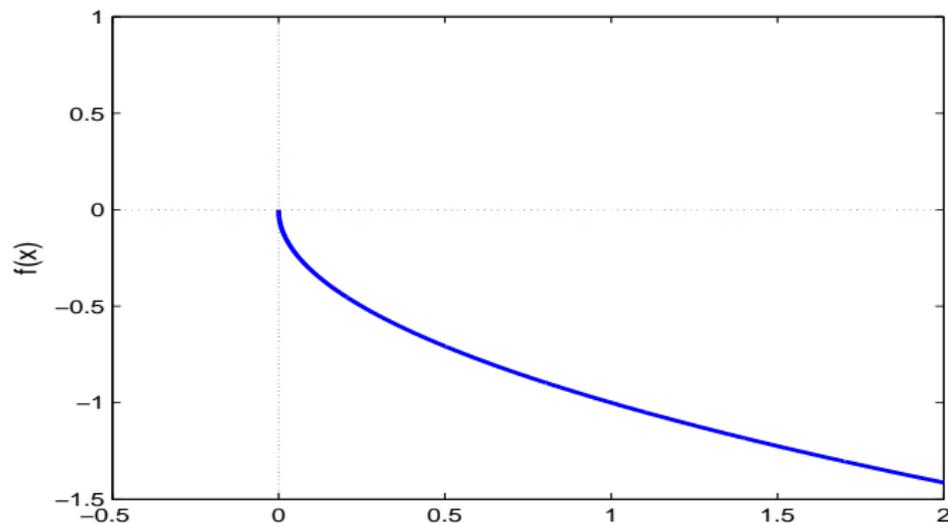
Subdifferentiability

- 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, and assume that $\tilde{\mathbf{x}} \in \operatorname{int}(\operatorname{dom}(f))$. Then $\partial f(\tilde{\mathbf{x}})$ is nonempty and bounded.

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 $X \subseteq \operatorname{int}(\operatorname{dom}(f))$ is nonempty and compact. Then $Y = \bigcup_{\mathbf{x} \in X} \partial f(\mathbf{x})$ is nonempty and bounded.

Nonemptiness of $\partial f(\mathbf{x})$ at the Relative Interior

- **Reminder:** Given a vector space $\mathbb{E}$, a set $S \subseteq \mathbb{E}$ is called affine if for any $\mathbf{x}, \mathbf{y} \in S, \lambda \in \mathbb{R}$, the inclusion $\lambda\mathbf{x} + (1 - \lambda)\mathbf{y} \in S$ holds.
- For a set $S \subseteq \mathbb{E}$, the affine hull of S, denoted by aff S, is the intersection of all affine sets containing S.
- Clearly, aff(S), is by itself an affine set, and it is the smallest affine set containing S (w.r.t. inclusion).
- The interior of a set S w.r.t. its affine hull, is called the relative interior of S:

$$\mathrm{ri}(S) = \{\mathbf{x} \in \mathrm{aff}(S) : \exists \varepsilon > 0 \, \mathrm{s.t.} \, B[\mathbf{x}, \varepsilon] \cap \mathrm{aff}\, S \subseteq S\}$$

Theorem. Let $C \subseteq \mathbb{E}$ be a convex set. Then $\mathrm{ri}(C)$ is nonempty.

Theorem. Let f be a proper extended real-valued convex function, and let $\mathbf{x} \in \mathrm{ri}(\mathrm{dom}\, f)$. Then $\partial f(\mathbf{x}) \neq \emptyset$.

Corollary. Let f be a proper extended real-valued convex function. Then there exists $\mathbf{x} \in \mathrm{dom}\, f$ for which $\partial f(\mathbf{x}) \neq \emptyset$.

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 $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$, 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 $\mathbf{x} \in \mathrm{int}(\mathrm{dom}\, f)$. 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 $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$. 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 $\mathbf{x} \in \text{dom}(f_1) \cap \text{dom}(f_2)$. 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 $\mathbf{x} \in \text{int}(\text{dom}(f_1)) \cap \text{int}(\text{dom}(f_2))$, then (**strong result**):
>
> $$\partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}) = \partial(f_1 + f_2)(\mathbf{x}).$$

Proof of (a).

- (a) Let $\mathbf{g} \in \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x})$.
- $\exists \mathbf{g}_1 \in \partial f_1(\mathbf{x}), \mathbf{g}_2 \in \partial f_2(\mathbf{x})$ such that $\mathbf{g} = \mathbf{g}_1 + \mathbf{g}_2$.
- $\forall \mathbf{y} \in \text{dom}(f_1 + f_2)$ $f_1(\mathbf{y}) \geq f_1(\mathbf{x}) + \langle \mathbf{g}_1, \mathbf{y} - \mathbf{x} \rangle, f_2(\mathbf{y}) \geq f_2(\mathbf{x}) + \langle \mathbf{g}_2, \mathbf{y} - \mathbf{x} \rangle$.
- Summing the two inequalities: $f_1(\mathbf{y}) + f_2(\mathbf{y}) \geq f_1(\mathbf{x}) + f_2(\mathbf{x}) + \langle \mathbf{g}_1 + \mathbf{g}_2, \mathbf{y} - \mathbf{x} \rangle$, implying that $\mathbf{g} = \mathbf{g}_1 + \mathbf{g}_2 \in \partial(f_1 + f_2)(\mathbf{x})$.

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

Example: $f(\mathbf{x}) = \|\mathbf{x}\|_1$.

- $f = \sum_{i=1}^{n} f_i$ where $f_i(\mathbf{x}) \equiv |x_i|$.
- $\partial f_i(\mathbf{x}) = \begin{cases} \{\text{sgn}(x_i)\mathbf{e}_i\}, & x_i \neq 0, \\ [-\mathbf{e}_i, \mathbf{e}_i], & x_i = 0. \end{cases}$
- By the sum rule of subdifferential calculus,

$$\partial f(\mathbf{x}) = \sum_{i=1}^{n} \partial f_i(\mathbf{x}) = \sum_{i \in I_{\neq}(\mathbf{x})} \text{sgn}(x_i)\mathbf{e}_i + \sum_{i \in I_0(\mathbf{x})} [-\mathbf{e}_i, \mathbf{e}_i],$$

where $I_{\neq}(\mathbf{x}) = \{i : x_i \neq 0\}$, $I_0(\mathbf{x}) = \{i : x_i = 0\}$.

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:

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

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

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

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

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

Proof - First Direction

- Let $\mathbf{x} \in \text{dom}(h)$ and assume that $\mathbf{g} \in \mathcal{A}^T \partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b})$.
- There exists $\mathbf{g} = \mathcal{A}^T(\mathbf{d})$, where $\mathbf{d} \in \partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b})$.
- Therefore, for any $\mathbf{y} \in \text{dom}(h)$ we have $\mathcal{A}(\mathbf{y}) + \mathbf{b} \in \text{dom}(f)$ and hence

$$f(\mathcal{A}(\mathbf{y}) + \mathbf{b}) \geq f(\mathcal{A}(\mathbf{x}) + \mathbf{b}) + \langle \mathbf{d}, \mathcal{A}(\mathbf{y}) + \mathbf{b} - \mathcal{A}(\mathbf{x}) - \mathbf{b} \rangle,$$

- $h(\mathbf{y}) \geq h(\mathbf{x}) + \langle \mathcal{A}^T(\mathbf{d}), \mathbf{y} - \mathbf{x} \rangle$ for all $\mathbf{y} \in \text{dom}(h)$
- Thus, $\mathbf{g} = \mathcal{A}^T(\mathbf{d}) \in \partial h(\mathbf{x})$, proving that $\mathcal{A}^T \partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b}) \subseteq \partial h(\mathbf{x})$.

Example: $\partial f(\mathbf{x})$, $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} + \mathbf{b}\|_1$

$\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{b} \in \mathbb{R}^m$.

- $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x} + \mathbf{b})$ with $g(\mathbf{y}) = \|\mathbf{y}\|_1$.
- $\partial f(\mathbf{x}) = \mathbf{A}^T \partial g(\mathbf{A}\mathbf{x} + \mathbf{b})$.
- Denote the ith row of $\mathbf{A}$ by $\mathbf{a}_i^T$ and define

$$
\begin{aligned}
I_{\neq}(\mathbf{x}) &= \{i : \mathbf{a}_i^T \mathbf{x} + b_i \neq 0\}, \\
I_0(\mathbf{x}) &= \{i : \mathbf{a}_i^T \mathbf{x} + b_i = 0\}.
\end{aligned}
$$

- In this terminology,
  $\partial g(\mathbf{A}\mathbf{x} + \mathbf{b}) = \sum_{i \in I_{\neq}(\mathbf{x})} \operatorname{sgn}(\mathbf{a}_i^T \mathbf{x} + b_i) \mathbf{e}_i + \sum_{i \in I_0(\mathbf{x})} [-\mathbf{e}_i, \mathbf{e}_i]$.

-
$$
\begin{aligned}
\partial f(\mathbf{x}) &= \mathbf{A}^T \partial g(\mathbf{A}\mathbf{x} + \mathbf{b}) \\
&= \sum_{i \in I_{\neq}(\mathbf{x})} \operatorname{sgn}(\mathbf{a}_i^T \mathbf{x} + b_i) \mathbf{A}^T \mathbf{e}_i + \sum_{i \in I_0(\mathbf{x})} [-\mathbf{A}^T \mathbf{e}_i, \mathbf{A}^T \mathbf{e}_i].
\end{aligned}
$$

Using the relation $\mathbf{A}^T \mathbf{e}_i = \mathbf{a}_i$,

$$
\boxed{\partial f(\mathbf{x}) = \sum_{i \in I_{\neq}(\mathbf{x})} \operatorname{sgn}(\mathbf{a}_i^T \mathbf{x} + b_i) \mathbf{a}_i + \sum_{i \in I_0(\mathbf{x})} [-\mathbf{a}_i, \mathbf{a}_i].}
$$

Example: $\partial f(\mathbf{x})$, $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} + \mathbf{b}\|_2$

$$\partial f(\mathbf{x}) = \mathbf{A}^T \partial g(\mathbf{A}\mathbf{x} + \mathbf{b}) = \begin{cases} \frac{\mathbf{A}^T(\mathbf{A}\mathbf{x} + \mathbf{b})}{\|\mathbf{A}\mathbf{x} + \mathbf{b}\|_2}, & \mathbf{A}\mathbf{x} + \mathbf{b} \neq \mathbf{0}, \\ \mathbf{A}^T B_{\|\cdot\|_2}[\mathbf{0}, 1], & \mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{0}. \end{cases}$$

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 non-decreasing 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

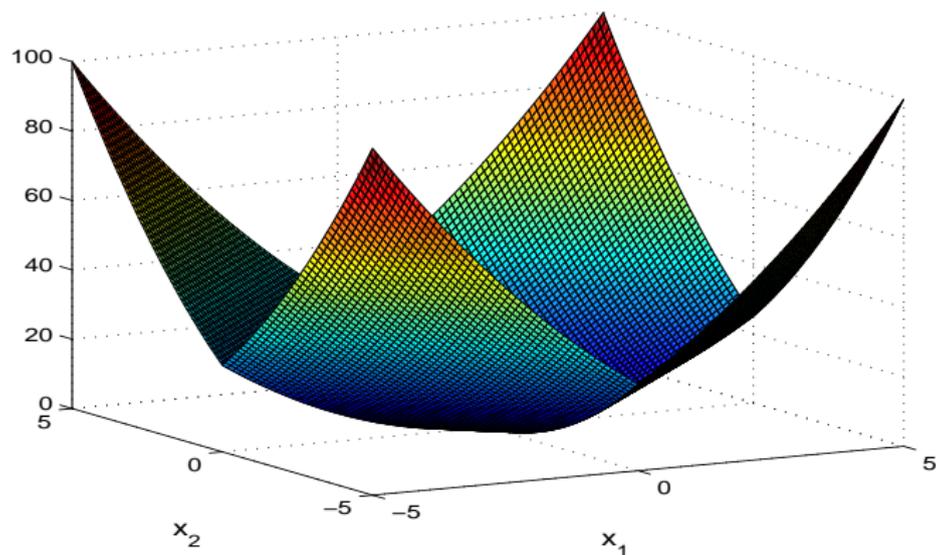$$\partial h(\mathbf{x}) = g'(f(\mathbf{x}))\partial f(\mathbf{x}).$$

Example

- $h(\mathbf{x}) = \|\mathbf{x}\|_1^2 \ (h : \mathbb{R}^n \to \mathbb{R})$

> $\partial h(\mathbf{x}) = 2\|\mathbf{x}\|_1 \{\mathbf{z} \in \mathbb{R}^n : z_i = \mathrm{sgn}(x_i), i \in I_{\neq}(\mathbf{x}), |z_j| \leq 1, j \in I_0(\mathbf{x})\},$
> where $I_{\neq}(\mathbf{x}) = \{i : x_i \neq 0\}, I_0(\mathbf{x}) = \{i : x_i = 0\}.$

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

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

where

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

Examples

- $f(\mathbf{x}) = \max\{x_1, x_2, \ldots, x_n\}$
- $f(\mathbf{x}) = \|\mathbf{x}\|_\infty$
- $f(\mathbf{x}) = \max_{i=1,2,\ldots,m}\{\mathbf{a}_i^T \mathbf{x} + b_i\}$ $(\mathbf{a}_i \in \mathbb{R}^n, b_i \in \mathbb{R})$
- $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} + \mathbf{b}\|_\infty$ $(\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{b} \in \mathbb{R}^m)$

Weak Maximum Rule of Subdifferential Calculus

Theorem. Let I be an arbitrary set, and suppose that any $i \in I$ is associated with a proper convex function $f_i : \mathbb{E} \to (-\infty, \infty]$. Let

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

Then for any $\mathbf{x} \in \text{dom}(f)$

$$\text{conv}\left(\cup_{i \in I(\mathbf{x})} \partial f_i(\mathbf{x})\right) \subseteq \partial f(\mathbf{x}),$$

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

Proof.

- Let $\mathbf{x} \in \text{dom}(f)$. Then for any $\mathbf{z} \in \text{dom}(f), i \in I(\mathbf{x})$ and $\mathbf{g} \in \partial f_i(\mathbf{x})$,

$$f(\mathbf{z}) \geq f_i(\mathbf{z}) \geq f_i(\mathbf{x}) + \langle \mathbf{g}, \mathbf{z} - \mathbf{x} \rangle = f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{z} - \mathbf{x} \rangle.$$

- Thus, $\mathbf{g} \in \partial f(\mathbf{x})$.
- $\partial f_i(\mathbf{x}) \subseteq \partial f(\mathbf{x})$.
- By the convexity of ∂f, $\text{conv}\left(\cup_{i \in I(\mathbf{x})} \partial f_i(\mathbf{x})\right) \subseteq \partial f(\mathbf{x})$.

Example

$f : \mathbb{R}^m \to \mathbb{R}$ given by $f(\mathbf{x}) = \lambda_{\max}(\mathcal{A}(\mathbf{x}))$, where

$$\mathcal{A}(\mathbf{x}) = \mathbf{A}_0 + \sum_{i=1}^{m} x_i \mathbf{A}_i \text{ for any } \mathbf{x} \in \mathbb{R}^m.$$

▶ $f(\mathbf{x}) = \max_{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y}\|_2 = 1} \mathbf{y}^T \mathcal{A}(\mathbf{x}) \mathbf{y}$

▶ $f_{\mathbf{y}}(\mathbf{x}) \equiv \mathbf{y}^T \mathcal{A}(\mathbf{x}) \mathbf{y} = \mathbf{y}^T \mathbf{A}_0 \mathbf{y} + \sum_{i=1}^{m} (\mathbf{y}^T \mathbf{A}_i \mathbf{y}) x_i$ is affine in $\mathbf{x}$, and in particular convex.

▶ The maximum in is attained at normalized eigenvectors corresponding to the maximum eigenvalue of $\mathcal{A}(\mathbf{x})$.

▶ By the weak maximum rule of subdifferential calculus, given a normalized eigenvector $\tilde{\mathbf{y}}$ of $\mathcal{A}(\mathbf{x})$ corresponding to the maximum eigenvalue

$$(\tilde{\mathbf{y}}^T \mathbf{A}_1 \tilde{\mathbf{y}}, \tilde{\mathbf{y}}^T \mathbf{A}_2 \tilde{\mathbf{y}}, \ldots, \tilde{\mathbf{y}}^T \mathbf{A}_m \tilde{\mathbf{y}})^T \in \partial f(\mathbf{x}),$$

Lipschitz Continuity and Boundedness of Subgradients

> **Theorem.** Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper convex extended real-valued function. Suppose that $X \subseteq \operatorname{int}(\operatorname{dom} f)$. Then the following two claims are equivalent:
>
> (a) $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in X$.
>
> (b) $\|\mathbf{g}\| \leq L$ for any $\mathbf{g} \in \partial f(\mathbf{x}), \mathbf{x} \in X$.

Proof. (will not be covered in class)

- Suppose that (b) is satisfied, and let $\mathbf{x}, \mathbf{y} \in X$.
- Let $\mathbf{g_x} \in \partial f(\mathbf{x}), \mathbf{g_y} \in \partial f(\mathbf{y})$.
- (a) follows by:

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \langle \mathbf{g_x}, \mathbf{x} - \mathbf{y} \rangle \leq \|\mathbf{g_x}\| \cdot \|\mathbf{x} - \mathbf{y}\| \leq L\|\mathbf{x} - \mathbf{y}\|,$$
$$f(\mathbf{y}) - f(\mathbf{x}) \leq \langle \mathbf{g_y}, \mathbf{y} - \mathbf{x} \rangle \leq \|\mathbf{g_y}\| \cdot \|\mathbf{x} - \mathbf{y}\| \leq L\|\mathbf{x} - \mathbf{y}\|,$$

- Now assume that (a) is satisfied. Take $\mathbf{x} \in X$ and $\mathbf{g} \in \partial f(\mathbf{x})$. We will show that $\|\mathbf{g}\| \leq L$.
- Define $\tilde{\mathbf{g}} \in \mathbb{E}$ that satisfies $\|\tilde{\mathbf{g}}\| = 1, \langle \tilde{\mathbf{g}}, \mathbf{g} \rangle = \|\mathbf{g}\|$.
- Take $\varepsilon > 0$ for which $\mathbf{x} + \varepsilon\tilde{\mathbf{g}} \in \operatorname{int}(\operatorname{dom} f)$.
- $\varepsilon\|\mathbf{g}\| = \langle \mathbf{g}, \varepsilon\tilde{\mathbf{g}} \rangle \leq f(\mathbf{x} + \varepsilon\tilde{\mathbf{g}}) - f(\mathbf{x}) \leq L\|\mathbf{x} + \varepsilon\tilde{\mathbf{g}} - \mathbf{x}\| = L\varepsilon.$

Fermat's Optimality Condition

Theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be an extended real-valued convex function. Then
$$\mathbf{x}^* \in \arg\min\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}$$
if and only if
$$\mathbf{0} \in \partial f(\mathbf{x}^*)$$

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

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

which is the the same as the inclusion $\mathbf{0} \in \partial f(\mathbf{x}^*)$.

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 $\left\| \sum_{i=1, i \neq j}^m \omega_i \frac{\mathbf{x}^* - \mathbf{a}_i}{\|\mathbf{x}^* - \mathbf{a}_i\|_2} \right\|_2 \leq \omega_j$.

[Sturm, 1884] [Weiszfeld, 1937]

Optimality Conditions of Convex Problems

> Theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be proper and convex, and let $C \subseteq \text{dom}(f)$ be a convex set. Then $\mathbf{x}^* \in C$ is an optimal solution of
>
> $$\min\{f(\mathbf{x}) : \mathbf{x} \in C\} \tag{2}$$
>
> iff there exists $\mathbf{g} \in \partial f(\mathbf{x}^*)$ for which $-\mathbf{g} \in N_C(\mathbf{x}^*)$.
> iff there exists $\mathbf{g} \in \partial f(\mathbf{x}^*)$ for which $\langle \mathbf{g}, \mathbf{x} - \mathbf{x}^* \rangle \geq 0$ for any $\mathbf{x} \in C$.

Proof.

- Problem (2) can be rewritten as $\min_{\mathbf{x} \in \mathbb{E}}\{f(\mathbf{x}) + \delta_C(\mathbf{x})\}$.
- $\text{ri}(\text{dom}(f)) \cap \text{ri}(C) = \text{ri}(C) \neq \emptyset$. Thus, by the sum rule of subdifferential calculus $\forall \mathbf{x} \in C \ \partial(f + \delta_C)(\mathbf{x}) = \partial f(\mathbf{x}) + \partial \delta_C(\mathbf{x}) = \partial f(\mathbf{x}) + N_C(\mathbf{x})$.
- By Fermat's optimality condition, $\mathbf{x}^* \in C$ is an optimal solution of (2) iff $\mathbf{0} \in \partial f(\mathbf{x}^*) + N_C(\mathbf{x}^*)$, that is, if and only if $(-\partial f(\mathbf{x}^*)) \cap N_C(\mathbf{x}^*) \neq \emptyset$.

Extension to 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 $\mathrm{dom}(g) \subseteq \mathrm{int}(\mathrm{dom}(f))$. Consider the problem

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

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

$$- \nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*). \tag{3}$$

(b) **(necessary and sufficient condition for convex problems)** Suppose that f is convex. If f is differentiable at $\mathbf{x}^* \in \mathrm{dom}(g)$, then $\mathbf{x}^*$ is a global optimal solution of (P) if and only if (3) is satisfied.

Proof

- Let $\mathbf{y} \in \operatorname{dom} g$.
- For any $\lambda \in (0, 1)$, define the point $\mathbf{x}_\lambda = (1 - \lambda)\mathbf{x}^* + \lambda \mathbf{y} \in \operatorname{dom} g$.
- For small enough λ

$$f(\mathbf{x}_\lambda) + g(\mathbf{x}_\lambda) \geq f(\mathbf{x}^*) + g(\mathbf{x}^*).$$

- $f((1 - \lambda)\mathbf{x}^* + \lambda \mathbf{y}) + g((1 - \lambda)\mathbf{x}^* + \lambda \mathbf{y}) \geq f(\mathbf{x}^*) + g(\mathbf{x}^*).$

Proof Contd.

- $f((1 - \lambda)\mathbf{x}^* + \lambda\mathbf{y}) + (1 - \lambda)g(\mathbf{x}^*) + \lambda g(\mathbf{y}) \geq f(\mathbf{x}^*) + g(\mathbf{x}^*)$,
- $\frac{f((1 - \lambda)\mathbf{x}^* + \lambda\mathbf{y}) - f(\mathbf{x}^*)}{\lambda} \geq g(\mathbf{x}^*) - g(\mathbf{y})$.
- Taking $\lambda \to 0^+$
- $f'(\mathbf{x}^*; \mathbf{y} - \mathbf{x}) \geq g(\mathbf{x}^*) - g(\mathbf{y})$.
- f is continuously differentiable, $f'(\mathbf{x}^*; \mathbf{y} - \mathbf{x}) = \langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x} \rangle$.
- $g(\mathbf{y}) \geq g(\mathbf{x}^*) + \langle -\nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle$,
- $-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*)$.
- If f is convex, we only need to show that if $-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*)$ is satisfied, then $\mathbf{x}^*$ is an optimal solution of (P).
- $g(\mathbf{y}) \geq g(\mathbf{x}^*) + \langle -\nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle$,
- convexity of f implies $f(\mathbf{y}) \geq f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle$.
- $f(\mathbf{y}) + g(\mathbf{y}) \geq f(\mathbf{x}) + g(\mathbf{x})$.

Stationarity in Composite Models

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

- $f : \mathbb{E} \to (-\infty, \infty]$ 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: $\min f(\mathbf{x}) + \lambda \|\mathbf{x}\|_1$ $(f : \mathbb{R}^n \to \mathbb{R})$

References

- ▶ 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).

Lecture 4 - 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). Why?

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

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

$$\delta_C^* = \sigma_C.$$

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 $\mathbf{x} \in \mathbb{E}$.

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

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

Properness of the Conjugate

> **Theorem.** Let $f : \mathbb{E} \to (-\infty, \infty]$ be a convex and proper function. Then f^* is proper.

Proof.

- Since f is proper, $\exists \hat{\mathbf{x}} \in \mathbb{E}$ such that $f(\hat{\mathbf{x}}) < \infty$.
- $f^*(\mathbf{y}) \geq \langle \mathbf{y}, \hat{\mathbf{x}} \rangle - f(\hat{\mathbf{x}}) > -\infty$,
- $\exists \mathbf{x} \in \mathrm{dom}\, f$ such that $\partial f(\mathbf{x}) \neq \emptyset$.
- Take $\mathbf{g} \in \partial f(\mathbf{x})$. Therefore, for any $\mathbf{z} \in \mathrm{dom}\, f$:

$$f(\mathbf{z}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{z} - \mathbf{x} \rangle.$$

- $f^*(\mathbf{g}) = \max_{\mathbf{x} \in \mathbb{E}} \{\langle \mathbf{g}, \mathbf{z} \rangle - f(\mathbf{z})\} \leq \langle \mathbf{g}, \mathbf{x} \rangle - f(\mathbf{x}) < \infty$,

Fenchel's Inequality

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

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

$f^{**} = f$

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

Proof (will not be covered in class).

- Suppose by contradiction that for some $\mathbf{x}$, we have $f^{**}(\mathbf{x}) < f(\mathbf{x})$.
- This means that $(\mathbf{x}, f^{**}(\mathbf{x})) \notin \mathrm{epi}(f)$.
- $\mathrm{epi}(f)$ is closed and convex.
- there exists $\mathbf{a} \in \mathbb{E}, b, c_1, c_2 \in \mathbb{R}$ such that

$$\langle \mathbf{a}, \mathbf{z} \rangle + bs \leq c_1 < c_2 \leq \langle \mathbf{a}, \mathbf{x} \rangle + bf^{**}(\mathbf{x}) \text{ for all } (\mathbf{z}, s) \in \mathrm{epi}(f).$$

- $\langle \mathbf{a}, \mathbf{z} - \mathbf{x} \rangle + b(s - f^{**}(\mathbf{x})) \leq c_1 - c_2 \equiv c < 0$ for all $(\mathbf{z}, s) \in \mathrm{epi}(f)$.
- $b \leq 0$

$f^{**} = f$ - Proof Contd. Case I: $b < 0$

▶ Take $\mathbf{y} = -\frac{\mathbf{a}}{b}$. Then

$$\langle \mathbf{y}, \mathbf{z} - \mathbf{x} \rangle - s + f^{**}(\mathbf{x}) \leq \frac{c}{-b} < 0.$$

▶ Choosing $s = f(\mathbf{z})$:

$$\langle \mathbf{y}, \mathbf{z} \rangle - f(\mathbf{z}) - \langle \mathbf{y}, \mathbf{x} \rangle + f^{**}(\mathbf{x}) \leq \frac{c}{-b} < 0.$$

▶ Taking the maximum over $\mathbf{z}$,

$$f^*(\mathbf{y}) - \langle \mathbf{y}, \mathbf{x} \rangle + f^{**}(\mathbf{x}) \leq \frac{c}{-b} < 0.$$

▶ Contradiction to Fenchel's inequality.

$f^{**} = f$ - Proof Contd. Case II: $b = 0$

▶ Take some $\hat{\mathbf{y}} \in \mathrm{dom}\, f^*$.

▶ Define

$$\begin{pmatrix} \hat{\mathbf{a}} \\ \hat{\mathbf{b}} \end{pmatrix} = \begin{pmatrix} a + \varepsilon \hat{\mathbf{y}} \\ -\varepsilon \end{pmatrix}.$$

▶ Then

$$\begin{aligned}
\langle \hat{\mathbf{a}}, \mathbf{z} - \mathbf{x} \rangle + \hat{b}(f(\mathbf{z}) - f^{**}(\mathbf{x})) &= c + \varepsilon[\langle \hat{\mathbf{y}}, \mathbf{z} \rangle - f(\mathbf{z}) + f^{**}(\mathbf{x}) - \langle \hat{\mathbf{y}}, \mathbf{x} \rangle] \\
&\leq c + \varepsilon[f^*(\hat{\mathbf{y}}) - \langle \hat{\mathbf{y}}, \mathbf{x} \rangle + f^{**}(\mathbf{x})] \equiv \hat{c}.
\end{aligned}$$

▶ Take $\varepsilon > 0$ small enough to make the righthand side of the above inequality to be negative.

▶ Define $\mathbf{y} = -\frac{\hat{\mathbf{a}}}{\hat{b}}$, and maximize w.r.t. $\mathbf{z}$:

$$f^*(\mathbf{y}) - \langle \mathbf{y}, \mathbf{x} \rangle + f^{**}(\mathbf{x}) \leq \frac{\hat{c}}{-\hat{b}} < 0,$$

▶ Contradiction to Fenchel's inequality.

Examples

- $f = \sigma_C$, where C is nonempty closed and convex. Then

$$f^* = \sigma_C^* = (\delta_C^*)^* = \delta_C^{**} = \delta_C.$$

- $f = \sigma_C$, where C is nonempty. Then

$$\sigma_C^* = \sigma_{\mathrm{cl(conv}(C))}^* = (\delta_{\mathrm{cl(conv}(C))}^*)^* = \delta_{\mathrm{cl(conv}(C))}^{**} = \delta_{\mathrm{cl(conv}(C))}.$$

- $f(\mathbf{x}) = \max\{x_1, x_2, \ldots, x_n\}$. Then $f^* = \delta_{\Delta_n}$. Why?

Simple Algebraic 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	$\sigma_C(\mathbf{x})$	C arbitrary
$\sigma_C(\mathbf{x})$	$\mathbb{R}^n$	$\delta_C(\mathbf{x})$	C closed, convex
$\|\mathbf{x}\|$	$\mathbb{R}^n$	$\delta_{B_{\|\cdot\|_*}[\mathbf{0},1]}$	$\|\cdot\|$ arbitrary norm
$-\sqrt{1 - \|\mathbf{x}\|^2}$	$B_{\|\cdot\|}[\mathbf{0},1]$	$\sqrt{\|\mathbf{y}\|_*^2 + 1}$	$\|\cdot\|$ arbitrary norm
$\frac{1}{p}|x|^p$	$\mathbb{R}$	$\frac{1}{q}|y|^q$	$p > 1, \frac{1}{p} + \frac{1}{q} = 1$
$\frac{1}{2}\|\mathbf{x}\|^2$	$\mathbb{R}^n$	$\frac{1}{2}\|\mathbf{y}\|_*^2$	$\|\cdot\|$ 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) $\mathbf{x} \in \partial f^*(\mathbf{y})$.

Proof.

- $\mathbf{y} \in \partial f(\mathbf{x})$ iff

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

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

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

- which combined with Fenchel's inequality is equivalent to $\langle \mathbf{x}, \mathbf{y} \rangle = f(\mathbf{x}) + f^*(\mathbf{y})$.

- If f is closed, then (i) is the same as $\langle \mathbf{x}, \mathbf{y} \rangle = g(\mathbf{y}) + g^*(\mathbf{x})$, where $g = f^*$, and thus it is equivalent to $\mathbf{x} \in \partial g(\mathbf{y}) = \partial f^*(\mathbf{y})$.

Conjugate Subgradient Theorem Contd.

- The equivalence $(i) \Leftrightarrow (ii)$ can be compactly written as

$$\partial f^*(\mathbf{y}) = \operatorname*{argmax}_{\mathbf{x}} \{\langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x})\}$$

- If f is closed, the equivalence $(i) \Leftrightarrow (iii)$ is the same as

$$\partial f(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y}} \{\langle \mathbf{x}, \mathbf{y} \rangle - f^*(\mathbf{y})\}$$

- If f is differentiable at a point $\mathbf{x}$, then

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

- In particular,

$$\partial f^*(\mathbf{0}) = \operatorname*{argmin}_{\mathbf{x}} f(\mathbf{x}).$$

- Example: $g(\mathbf{x}) = \|\mathbf{x}\|$ (arbitrary norm). What is $\partial g(\mathbf{0})$?

Fenchel's Duality Theorem

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

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:

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

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

References

- 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).

Lecture 5 - The Proximal Operator

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

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

Examples

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

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

  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

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

- 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 $\text{prox}_f(\mathbf{x})$ 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

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

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 $\text{dom}(f)$ is convex and the following inequality holds for any $\mathbf{x}, \mathbf{y} \in \text{dom}(f)$ 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 given in the theorem is relevant only for Euclidean spaces.

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 $\lambda_{\min}(\mathbf{A})$.

Strongly Convex+Convex=Strongly Convex

> **Lemma.** Let $f : \mathbb{E} \to (-\infty, \infty]$ be a σ-strongly convex function ($\sigma > 0$), and let $g : \mathbb{E} \to (-\infty, \infty]$ be convex. Then $f + g$ is σ-strongly convex.

Proof.

- Since f and g are convex, both $\text{dom}(f)$ and $\text{dom}(g)$ are convex sets and hence also $\text{dom}(f + g) = \text{dom}(f) \cap \text{dom}(g)$ is a convex set.
- Let $\mathbf{x}, \mathbf{y} \in \text{dom}(f) \cap \text{dom}(g)$ and $\lambda \in [0, 1]$. By the σ-strong convexity of f,

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

- Since g is convex, $g(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda g(\mathbf{x}) + (1 - \lambda)g(\mathbf{y})$.
- Adding the two inequalities,

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

- Hence, $f + g$ is σ-strongly convex.

Consequence: $\frac{1}{2}\|\mathbf{x}\|^2 + \delta_C(\mathbf{x})$ is 1-strongly convex whenever $C \subseteq \mathbb{E}$ is nonempty and convex.

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 $\mathbf{x} \in \mathrm{dom}(\partial f), \mathbf{y} \in \mathrm{dom}(f)$ and $\mathbf{g} \in \partial f(\mathbf{x})$.

(iii)
$$\langle \mathbf{g_x} - \mathbf{g_y}, \mathbf{x} - \mathbf{y} \rangle \geq \sigma \|\mathbf{x} - \mathbf{y}\|^2$$

for any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(\partial f)$ and $\mathbf{g_x} \in \partial f(\mathbf{x}), \mathbf{g_y} \in \partial f(\mathbf{y})$.

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 $\mathbf{x} \in \mathrm{dom}(f)$, where $\mathbf{x}^*$ is the unique minimizer of f.

Proof.

- Take $\mathbf{x}^0 \in \mathrm{dom}(\partial f)$ and $\mathbf{g} \in \partial f(\mathbf{x}^0)$. Then

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) + \langle \mathbf{g}, \mathbf{x} - \mathbf{x}_0 \rangle + \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}_0\|^2 \text{ for all } \mathbf{x} \in \mathbb{E}.$$

- Rearranging terms,

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) - \frac{1}{2\sigma}\|\mathbf{g}\|^2 + \frac{\sigma}{2}\left\|\mathbf{x} - \left(\mathbf{x}_0 - \frac{1}{\sigma}\mathbf{g}\right)\right\|^2 \text{ for any } \mathbf{x} \in \mathbb{E}.$$

- Thus, $\mathrm{Lev}(f, f(\mathbf{x}_0)) \subseteq B_{\|\cdot\|}\left[\mathbf{x}_0 - \frac{1}{\sigma}\mathbf{g}, \frac{1}{\sigma}\|\mathbf{g}\|\right]$.
- $\mathrm{Lev}(f, f(\mathbf{x}_0))$ is compact $\Rightarrow$+closedness of f $\Rightarrow$ a minimizer exists.

Proof Contd.

- To show uniqueness, if $\tilde{\mathbf{x}}$ and $\hat{\mathbf{x}}$ are minimizers of f. Then $f(\tilde{\mathbf{x}}) = f(\hat{\mathbf{x}}) = f_{\text{opt}}$. Then

$$f_{\text{opt}} \leq f\left(\frac{1}{2}\tilde{\mathbf{x}} + \frac{1}{2}\hat{\mathbf{x}}\right) \leq \frac{1}{2}f(\tilde{\mathbf{x}}) + \frac{1}{2}f(\hat{\mathbf{x}}) - \frac{\sigma}{8}\|\tilde{\mathbf{x}} - \hat{\mathbf{x}}\|^2 = f_{\text{opt}} - \frac{\sigma}{8}\|\tilde{\mathbf{x}} - \bar{\mathbf{x}}\|^2.$$

  showing that $\tilde{\mathbf{x}} = \hat{\mathbf{x}}$ (proving part (a)).

- Let $\mathbf{x}^*$ be the unique minimizer of f.

- By Fermat's optimality condition $\mathbf{0} \in \partial f(\mathbf{x}^*)$, and hence for all $\mathbf{x} \in \mathbb{E}$,

$$f(\mathbf{x}) - f(\mathbf{x}^*) \geq \langle \mathbf{0}, \mathbf{x} - \mathbf{x}^* \rangle + \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^*\|^2 = \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^*\|^2$$

Back to 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 $\mathbf{x} \in \mathbb{E}$,

$$\operatorname{prox}_f(\mathbf{x}) = \underset{\mathbf{u} \in \mathbb{E}}{\operatorname{argmin}} \, \tilde{f}(\mathbf{u}, \mathbf{x}), \qquad (1)$$

  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 (1).

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.

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

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

▶ 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, $\mathcal{A} \circ \mathcal{A}^T = \alpha I$, $\alpha > 0$
$g(\|\mathbf{x}\|)$	$\mathrm{prox}_g(\|\mathbf{x}\|)\frac{\mathbf{x}}{\|\mathbf{x}\|}, \qquad \mathbf{x} \neq \mathbf{0}$ $\{\mathbf{u} : \|\mathbf{u}\| = \mathrm{prox}_g(0)\}, \quad \mathbf{x} = \mathbf{0}$	g proper closed convex, $\mathrm{dom}(g) \subseteq [0, \infty)$

Proof of One Property

> Let $g : \mathbb{E} \to (-\infty, \infty]$ be proper, and let $\lambda \neq 0$. Define $f(\mathbf{x}) = \lambda g(\mathbf{x}/\lambda)$. Then $\mathrm{prox}_f(\mathbf{x}) = \lambda \mathrm{prox}_{g/\lambda}(\mathbf{x}/\lambda)$

Proof.

▶ Note that

$$\mathrm{prox}_f(\mathbf{x}) = \underset{\mathbf{u}}{\mathrm{argmin}} \left\{ f(\mathbf{u}) + \frac{1}{2} \|\mathbf{u} - \mathbf{x}\|^2 \right\} = \underset{\mathbf{u}}{\mathrm{argmin}} \left\{ \lambda g\left(\frac{\mathbf{u}}{\lambda}\right) + \frac{1}{2} \|\mathbf{u} - \mathbf{x}\|^2 \right\}.$$

▶ Making the change of variables $\mathbf{z} = \frac{\mathbf{u}}{\lambda}$, we can continue to write

$$
\begin{aligned}
\mathrm{prox}_f(\mathbf{x}) &= \lambda \underset{\mathbf{z}}{\mathrm{argmin}} \left\{ \lambda g(\mathbf{z}) + \frac{1}{2} \|\lambda \mathbf{z} - \mathbf{x}\|^2 \right\} \\
&= \lambda \underset{\mathbf{z}}{\mathrm{argmin}} \left\{ \lambda^2 \left[\frac{g(\mathbf{z})}{\lambda} + \frac{1}{2} \left\| \mathbf{z} - \frac{\mathbf{x}}{\lambda} \right\|^2 \right] \right\} \\
&= \lambda \underset{\mathbf{z}}{\mathrm{argmin}} \left\{ \frac{g(\mathbf{z})}{\lambda} + \frac{1}{2} \left\| \mathbf{z} - \frac{\mathbf{x}}{\lambda} \right\|^2 \right\} \\
&= \lambda \mathrm{prox}_{g/\lambda}(\mathbf{x}/\lambda).
\end{aligned}
$$

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, $\lambda > 0$
$\lambda\|\mathbf{x}\|_1$	$\mathbb{R}^n$	$[\|\mathbf{x}\| - \lambda\mathbf{e}]_+ \circ \mathrm{sgn}(\mathbf{x})$	$\lambda > 0$
$-\lambda \sum_{j=1}^n \log x_j$	$\mathbb{R}_{++}^n$	$\left(\frac{x_j + \sqrt{x_j^2 + 4\lambda}}{2}\right)_{j=1}^n$	$\lambda > 0$
$\delta_C(\mathbf{x})$	$\mathbb{E}$	$P_C(\mathbf{x})$	$C \subseteq \mathbb{E}$
$\lambda\sigma_C(\mathbf{x})$	$\mathbb{E}$	$\mathbf{x} - \lambda P_C(\mathbf{x}/\lambda)$	C closed and convex
$\lambda\|\mathbf{x}\|$	$\mathbb{E}$	$\mathbf{x} - \lambda P_{B_{\|\cdot\|_*}[0,1]}(\mathbf{x}/\lambda)$	arbitrary norm
$\lambda \max\{x_1, x_2, \ldots, x_n\}$	$\mathbb{R}^n$	$\mathbf{x} - \mathrm{prox}_{\Delta_n}(\mathbf{x}/\lambda)$	$\lambda > 0$
$\lambda d_C(\mathbf{x})$	$\mathbb{E}$	$\mathbf{x} + \min\left\{\frac{\lambda}{d_C(\mathbf{x})}, 1\right\}(P_C(\mathbf{x}) - \mathbf{x})$	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 $\varphi(t) = \lambda|t|$.

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

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

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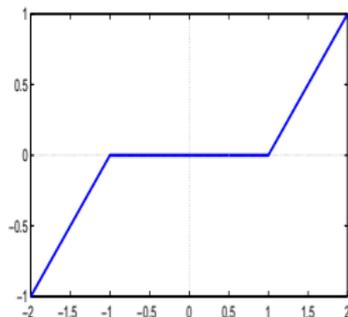
- By the separability of the l_1-norm, $\operatorname{prox}_g(\mathbf{x}) = (\mathcal{T}_\lambda(x_j))_{j=1}^n$. We expend the definition of the soft thresholding operator and write

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

Prox of $|\mathbf{a}^T\mathbf{x}|$

- $f(\mathbf{x}) = |\mathbf{a}^T\mathbf{x}|$ $(\mathbf{0} \neq \mathbf{a} \in \mathbb{R}^n)$.
- $f(\mathbf{x}) = g(\mathbf{a}^T\mathbf{x})$, where $g(t) = |t|$.
- $\operatorname{prox}_{\lambda g} = \mathcal{T}_\lambda$
- $\operatorname{prox}_f(\mathbf{x}) = \mathbf{x} + \frac{1}{\|\mathbf{a}\|^2}(\mathcal{T}_{\|\mathbf{a}\|^2}(\mathbf{a}^T\mathbf{x}) - \mathbf{a}^T\mathbf{x})\mathbf{a}$.

The Second Prox Theorem

> Theorem Let $g : \mathbb{E} \to (-\infty, \infty]$ be a proper, closed and convex function. Then
> (i) $\mathbf{u} = \text{prox}_g(\mathbf{x})$.
> (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 $\mathbf{x} = \text{prox}_f(\mathbf{x})$

Firm Nonexpansivity of the Prox Operator

> Theorem. For any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$
> (i) $\langle \mathbf{x} - \mathbf{y}, \operatorname{prox}_h(\mathbf{x}) - \operatorname{prox}_h(\mathbf{y}) \rangle \geq \|\operatorname{prox}_h(\mathbf{x}) - \operatorname{prox}_h(\mathbf{y})\|^2$.
> (ii) $\|\operatorname{prox}_h(\mathbf{x}) - \operatorname{prox}_h(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|$.

Proof.

- Denote $\mathbf{u} = \operatorname{prox}_h(\mathbf{x}), \mathbf{v} = \operatorname{prox}_h(\mathbf{y})$.
- $\mathbf{x} - \mathbf{u} \in \partial h(\mathbf{u}), \mathbf{y} - \mathbf{v} \in \partial h(\mathbf{v})$.
- 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 $\mathbf{x} \in \mathbb{E}$
>
> $$\operatorname{prox}_f(\mathbf{x}) + \operatorname{prox}_{f^*}(\mathbf{x}) = \mathbf{x}.$$

Proof.

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

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

A direct consequence (extended Moreau decomposition)

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

Prox of Support Functions

> Let C be a nonempty closed and convex set, and let $\lambda > 0$. Then
> $$\operatorname{prox}_{\lambda \sigma_C}(\mathbf{x}) = \mathbf{x} - \lambda P_C(\mathbf{x}/\lambda).$$

Proof. By the extended Moreau decomposition formula

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

Examples:

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

References

▶ 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).

Preliminaries to Lecture 6 – Smoothness

> **Definition.** Let $L \geq 0$. A function $f : \mathbb{E} \to (-\infty, \infty]$ is said to be *L-smooth* over a set $D \subseteq \mathrm{int}(\mathrm{dom}(f))$ 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) $f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{E}$.

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

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

(v) $f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \geq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}) - \frac{L}{2}\lambda(1-\lambda)\|\mathbf{x} - \mathbf{y}\|^2$ 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.

Lecture 6 - 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.

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

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

Proximal Gradient Method

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

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

(a) pick $L_k > 0$.

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

▶ The general update step can be written as $\mathbf{x}^{k+1} = T_{L_k}^{f,g}(\mathbf{x}^k)$

▶ $T_L^{f,g} : \text{int}(\text{dom}(f)) \to \mathbb{E}$ is the prox-grad operator defined by

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

▶ 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 $\mathbf{x} \in \text{dom}(f)$ and $L \in \left(\frac{L_f}{2}, \infty\right)$

$$F(\mathbf{x}) - F(T_L(\mathbf{x})) \geq \frac{L - \frac{L_f}{2}}{L^2} \left\| G_L^{f,g}(\mathbf{x}) \right\|^2, \tag{1}$$

where $G_L^{f,g} : \text{int}(\text{dom}(f)) \to \mathbb{E}$ 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
$$f(\mathbf{x}^+) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}^+ - \mathbf{x} \rangle + \frac{L_f}{2} \|\mathbf{x} - \mathbf{x}^+\|^2. \tag{2}$$

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

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

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

▶ which combined with (2) yields

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

The Gradient Mapping

- **Definition.** Suppose that f and g satisfy properties (A) and (B). Then 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 $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$:

$$F(\mathbf{x}) - F(T_{L_f}(\mathbf{x})) \geq \frac{1}{2L_f} \|G_{L_f}(\mathbf{x})\|^2.$$

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 $\mathbf{x} \in \text{int}(\text{dom}(f))$, where $g_0(\mathbf{x}) \equiv 0$.
>
> (b) For $\mathbf{x}^* \in \text{int}(\text{dom}(f))$, $G_L^{f,g}(\mathbf{x}^*) = \mathbf{0}$ iff $\mathbf{x}^*$ is a stationary point

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

(b) $G_L^{f,g}(\mathbf{x}^*) = \mathbf{0}$ iff $\mathbf{x}^* = \text{prox}_{\frac{1}{L}g}\left(\mathbf{x}^* - \frac{1}{L}\nabla f(\mathbf{x}^*)\right)$. 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 $\mathbf{x} \in \text{dom}(f)$ 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.** $\|G_L(\mathbf{x}) - G_L(\mathbf{y})\| \leq (2L + L_f)\|\mathbf{x} - \mathbf{y}\|$.

Proof of Lipschitz cont.

$$
\begin{aligned}
\|G_L(\mathbf{x}) - G_L(\mathbf{y})\| &= L \left\| \mathbf{x} - \text{prox}_{\frac{1}{L}g}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right) - \mathbf{y} + \text{prox}_{\frac{1}{L}g}\left(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\right) \right\| \\
&\leq L\|\mathbf{x} - \mathbf{y}\| + L \left\| \text{prox}_{\frac{1}{L}g}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right) - \text{prox}_{\frac{1}{L}g}\left(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\right) \right\| \\
&\leq L\|\mathbf{x} - \mathbf{y}\| + L \left\| \left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right) - \left(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\right) \right\| \\
&\leq 2L\|\mathbf{x} - \mathbf{y}\| + \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \\
&\leq (2L + L_f)\|\mathbf{x} - \mathbf{y}\|.
\end{aligned}
$$

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, $F(\mathbf{x}^{k+1}) < F(\mathbf{x}^k)$ unless $\mathbf{x}^k$ is a stationary point of (P).
>
> (b) $G_d(\mathbf{x}^k) \to \mathbf{0}$ as $k \to \infty$.
>
> (c) $\min_{n=0,1,\ldots,k} \|G_d(\mathbf{x}^n)\| \leq \frac{\sqrt{F(\mathbf{x}^0) - F_{\mathrm{opt}}}}{\sqrt{M(k+1)}}$.
>
> (d) All limit points of the sequence $\{\mathbf{x}^k\}_{k \geq 0}$ are stationary points of problem (P).

Proof.

- (a) follows by the inequality $F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq M\|G_d(\mathbf{x}^k)\|^2$.

- (b). Since the sequence $\{F(\mathbf{x}^k)\}_{k \geq 0}$ is nonincreasing and bounded below, it converges. Thus, in particular $F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \to 0$ as $k \to \infty$, implying that $\|G_d(\mathbf{x}^k)\| \to 0$.

Proof Contd.

- (c). Summing the inequality

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

  over $n = 0, 1, \ldots, k$, we obtain

$$F(\mathbf{x}^0) - F(\mathbf{x}^{k+1}) \geq M\sum_{n=0}^{k}\|G_d(\mathbf{x}^n)\|^2 \geq M(k+1)\min_{n=0,1,\ldots,k}\|G_d(\mathbf{x}^n)\|^2.$$

  Using the fact that $F(\mathbf{x}^{k+1}) \geq F_{\mathrm{opt}}$, (c) follows.

- (d). Let $\bar{\mathbf{x}}$ be a limit point of $\{\mathbf{x}^k\}_{k\geq 0}$. Then there exists a subsequence $\{\mathbf{x}^{k_j}\}_{j\geq 0}$ converging to $\bar{\mathbf{x}}$. For any $j \geq 0$,

$$\|G_d(\bar{\mathbf{x}})\| \leq \|G_d(\mathbf{x}^{k_j}) - G_d(\bar{\mathbf{x}})\| + \|G_d(\mathbf{x}^{k_j})\| \leq (2d + L_f)\|\mathbf{x}^{k_j} - \bar{\mathbf{x}}\| + \|G_d(\mathbf{x}^{k_j})\|,$$

  Since the RHS to 0 as $j \to \infty$, it follows that $G_d(\bar{\mathbf{x}}) = \mathbf{0}$, meaning that $\bar{\mathbf{x}}$ is a stationary point of problem (P).

The Fundamental Prox-Grad Inequality

> **Theorem.** For any $\mathbf{x} \in \mathbb{E}$ and $\mathbf{y} \in \text{int}(\text{dom}(f))$ satisfying
> $$f(T_L(\mathbf{y})) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), T_L(\mathbf{y}) - \mathbf{y} \rangle + \frac{L}{2} \| T_L(\mathbf{y}) - \mathbf{y} \|^2, \qquad (3)$$
>
> it holds that
> $$F(\mathbf{x}) - F(T_L(\mathbf{y})) \geq \frac{L}{2} \| \mathbf{x} - T_L(\mathbf{y}) \|^2 - \frac{L}{2} \| \mathbf{x} - \mathbf{y} \|^2 + \ell_f(\mathbf{x}, \mathbf{y}), \qquad (4)$$
>
> 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 $\mathbf{y}^+ = \text{prox}_{\frac{1}{L}g}\left(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\right)$, by the second prox theorem it follows that

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

- Therefore,

$$
\begin{aligned}
g(\mathbf{x}) \;\geq\;& g(\mathbf{y}^+) + L\langle \mathbf{y} - \mathbf{y}^+, \mathbf{x} - \mathbf{y}^+ \rangle + \langle \nabla f(\mathbf{y}), \mathbf{y}^+ - \mathbf{x} \rangle \\
=\;& g(\mathbf{y}^+) + L\langle \mathbf{y} - \mathbf{y}^+, \mathbf{x} - \mathbf{y}^+ \rangle \\
& + \langle \nabla f(\mathbf{y}), \mathbf{y}^+ - \mathbf{y} \rangle + \langle \nabla f(\mathbf{y}), \mathbf{y} - \mathbf{x} \rangle
\end{aligned} \qquad (5)
$$

Proof Contd.

- By (3), $f(\mathbf{y}^+) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{y}^+ - \mathbf{y} \rangle + \frac{L}{2}\|\mathbf{y}^+ - \mathbf{y}\|^2$
- Hence, $\langle \nabla f(\mathbf{y}), \mathbf{y}^+ - \mathbf{y} \rangle \geq f(\mathbf{y}^+) - f(\mathbf{y}) - \frac{L}{2}\|\mathbf{y}^+ - \mathbf{y}\|^2$,
- which combined with (5) yields

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

- 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

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

Sufficient Decrease Lemma - 2nd Version

Corollary. For any $\mathbf{x} \in \operatorname{int}(\operatorname{dom}(f))$ 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 Nonconvex 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

- $\beta L_f \leq L_k \leq \alpha L_f$, 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

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

or

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

$O(1/k)$ 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 $k \geq 0$,
>
> $$F(\mathbf{x}^k) - F_{\mathrm{opt}} \leq \frac{\alpha L_f \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2k},$$
>
> 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 $L_n \leq \alpha L_f$, we obtain

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

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

- By the monotonicity of $\{F(\mathbf{x}^n)\}_{n \geq 0}$,

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

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

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 (4),

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

▶ 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 a bounded sequence, 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.

Proof.

▶ Since $\{\mathbf{x}^k\}_{k\geq 0}$ is bounded, it has limit points. We need to show that it has a *unique* limit point.

▶ Let $\tilde{\mathbf{x}}$ be a limit point. There exists a subsequence $\{\mathbf{x}^{k_j}\}_{j\geq 0}$ such that $\mathbf{x}^{k_j} \to \tilde{\mathbf{x}}$.

▶ Since $\tilde{\mathbf{x}} \in D \subseteq S$, then $\|\mathbf{x}^{k+1} - \tilde{\mathbf{x}}\| \leq \|\mathbf{x}^k - \tilde{\mathbf{x}}\|$.

▶ Thus, $\{\|\mathbf{x}^k - \tilde{\mathbf{x}}\|\}_{k\geq 0}$ is nonincreasing and bounded below and hence convergent.

▶ Since $\|\mathbf{x}^{k_j} - \tilde{\mathbf{x}}\| \to 0$ as $j \to \infty$, it follows that the whole sequence $\{\|\mathbf{x}^k - \tilde{\mathbf{x}}\|\}_{k\geq 0}$ converges to zero.

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).

Proof.

- ▶ The sequence is Fejér monotone w.r.t. X^*.
- ▶ Therefore, it is enough to show that any limit point of the sequence $\{\mathbf{x}^k\}_{k\geq 0}$ is necessarily in X^*.
- ▶ Let $\tilde{\mathbf{x}}$ be a limit point. Then there exists a subsequence $\{\mathbf{x}^{k_j}\}_{j\geq 0}$ converging to $\tilde{\mathbf{x}}$.
- ▶ $F(\mathbf{x}^{k_j}) \to F_{\mathrm{opt}}$ as $j \to \infty$.
- ▶ F is lower semicontinuous and hence $F(\tilde{\mathbf{x}}) \leq \lim_{j\to\infty} F(\mathbf{x}^{k_j}) = F_{\mathrm{opt}}$, implying that $\tilde{\mathbf{x}} \in X^*$.

Iteration Complexity of Algorithms

- An ε-optimal solution of problem (P) is a vector $\bar{\mathbf{x}} \in \mathrm{dom}(g)$ satisfying $F(\bar{\mathbf{x}}) - F_{\mathrm{opt}} \leq \varepsilon$.
- 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 $F(\mathbf{x}^k) - F_{\mathrm{opt}} \leq \varepsilon$
- Recall: $F(\mathbf{x}^k) - F_{\mathrm{opt}} \leq \frac{\alpha L_f \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2k}$.

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

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.

Proof

- By the sufficient decrease lemma
  $F(\mathbf{x}^n) - F(\mathbf{x}^{n+1}) = F(\mathbf{x}^n) - F(T_{L_n}(\mathbf{x}^n)) \geq \frac{1}{2L_n} \|G_{L_n}(\mathbf{x}^n)\|^2$.

- Since $\beta L_f \leq L_n \leq \alpha L_f$ and by the monotonicity of the norm of the gradient mapping,

$$\frac{1}{2L_n} \|G_{L_n}(\mathbf{x}^n)\|^2 = \frac{L_n}{2} \frac{\|G_{L_n}(\mathbf{x}^n)\|^2}{L_n^2} \geq \frac{\beta L_f}{2} \frac{\|G_{\alpha L_f}(\mathbf{x}^n)\|^2}{\alpha^2 L_f^2} = \frac{\beta}{2\alpha^2 L_f} \|G_{\alpha L_f}(\mathbf{x}^n)\|^2$$

- $F(\mathbf{x}^n) - F_{\mathrm{opt}} \geq F(\mathbf{x}^{n+1}) - F_{\mathrm{opt}} + \frac{\beta}{2\alpha^2 L_f} \|G_{\alpha L_f}(\mathbf{x}^n)\|^2$.

- Summing over $n = p, p+1, \ldots, 2p-1$ yields
  $F(\mathbf{x}^p) - F_{\mathrm{opt}} \geq F(\mathbf{x}^{2p}) - F_{\mathrm{opt}} + \frac{\beta}{2\alpha^2 L_f} \sum_{n=p}^{2p-1} \|G_{\alpha L_f}(\mathbf{x}^n)\|^2$.

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

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

- $\min_{n=0,1,\ldots,k} \|G_{\alpha L_f}(\mathbf{x}^n)\|^2 \leq \frac{\alpha^3 L_f^2 \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{\beta \max\{(k/2)^2, ((k+1)/2)^2\}} = \frac{4\alpha^3 L_f^2 \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{\beta (k+1)^2}$.

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

Proof

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

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

- $\ell_f(\mathbf{x}^*, \mathbf{x}^k) = f(\mathbf{x}^*) - f(\mathbf{x}^k) - \langle \nabla f(\mathbf{x}^k), \mathbf{x}^* - \mathbf{x}^k \rangle \geq \frac{\sigma}{2}\|\mathbf{x}^k - \mathbf{x}^*\|^2$.
- $F(\mathbf{x}^*) - F(\mathbf{x}^{k+1}) \geq \frac{L_k}{2}\|\mathbf{x}^* - \mathbf{x}^{k+1}\|^2 - \frac{L_k - \sigma}{2}\|\mathbf{x}^* - \mathbf{x}^k\|^2$.
- Since $F(\mathbf{x}^*) - F(\mathbf{x}^{k+1}) \leq 0$ $L_k \leq \alpha L_f$,

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

establishing part (a). Part (b) follows immediately by (a).

- (c) follows by

$$\begin{aligned}
F(\mathbf{x}^{k+1}) - F_{\text{opt}} &\leq \frac{L_k - \sigma}{2}\|\mathbf{x}^k - \mathbf{x}^*\|^2 - \frac{L_k}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 \\
&\leq \frac{\alpha L_f}{2}\left(1 - \frac{\sigma}{\alpha L_f}\right)\|\mathbf{x}^k - \mathbf{x}^*\|^2 \\
&\leq \frac{\alpha L_f}{2}\left(1 - \frac{\sigma}{\alpha L_f}\right)^{k+1}\|\mathbf{x}^0 - \mathbf{x}^*\|^2,
\end{aligned}$$

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 $F(\mathbf{x}^k) - F_{\mathrm{opt}} \leq \varepsilon$, where R is an upper bound on $\|\mathbf{x}^0 - \mathbf{x}^*\|$ for some $\mathbf{x}^* \in X^*$ 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:
$$\|\mathbf{y}\|_* \equiv \max_{\mathbf{x}} \{\langle \mathbf{y}, \mathbf{x} \rangle : \|\mathbf{x}\| \leq 1\}, \quad \mathbf{y} \in \mathbb{E}^*.$$

- 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

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

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

(a) If $\mathbf{a} \ne \mathbf{0}$, 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: $\mathbb{E} = \mathbb{R}^n$, $\mathbf{a} \ne \mathbf{0}$,

▶ $\|\cdot\| = \|\cdot\|_2$ $\Lambda_{\mathbf{a}} = \left\{ \frac{\mathbf{a}}{\|\mathbf{a}\|_2} \right\}$.

▶ $\|\cdot\| = \|\cdot\|_1$, $\Lambda_{\mathbf{a}} = \left\{ \sum_{i \in I(\mathbf{a})} \lambda_i \operatorname{sgn}(a_i)\mathbf{e}_i : \sum_{i \in I(\mathbf{a})} \lambda_i = 1, \lambda_j \ge 0, j \in I(\mathbf{a}) \right\}$,
where $I(\mathbf{a}) = \underset{i=1,2,\dots,n}{\operatorname{argmax}} |a_i|$.

▶ $\|\cdot\| = \|\cdot\|_\infty$. $\Lambda_{\mathbf{a}} = \{\mathbf{z} \in \mathbb{R}^n : z_i = \operatorname{sgn}(a_i), i \in I_{\ne}(\mathbf{a}), |z_j| \le 1, j \in I_0(\mathbf{a})\}$,
where
$I_{\ne}(\mathbf{a}) = \{i \in \{1, 2, \dots, n\} : a_i \ne 0\}, I_0(\mathbf{a}) = \{i \in \{1, 2, \dots, n\} : a_i = 0\}$.

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_{opt}, then
$\min_{n=0,1,\ldots,k} \|\nabla f(\mathbf{x}^n)\|_* \leq \frac{\sqrt{2L_f}\sqrt{f(\mathbf{x}^0) - f_{\mathrm{opt}}}}{\sqrt{k+1}}$.

(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

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

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_{opt}.
- There exists $R > 0$ s.t. $\max_{\mathbf{x}, \mathbf{x}^*} \{\|\mathbf{x}^* - \mathbf{x}\| : f(\mathbf{x}) \leq f(\mathbf{x}^0), \mathbf{x}^* \in X^*\} \leq R$.

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

Proof.

- By the gradient inequality,

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

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

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

For any $k \geq 1$,
$$f(\mathbf{x}^k) - f_{\mathrm{opt}} \leq \frac{2L_f R^2}{k}$$

Proof.

- Define $a_k = f(\mathbf{x}^k) - f_{\mathrm{opt}}$
- 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 $i_k \in \underset{i}{\operatorname{argmax}} \left| \dfrac{\partial f(\mathbf{x}^k)}{\partial x_i} \right|$;
  - $\mathbf{x}^{k+1} = \mathbf{x}^k - \dfrac{\|\nabla f(\mathbf{x}^k)\|_\infty}{L_f} \operatorname{sgn}\left(\dfrac{\partial f(\mathbf{x}^k)}{\partial x_{i_k}} \right) \mathbf{e}_{i_k}$.

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).
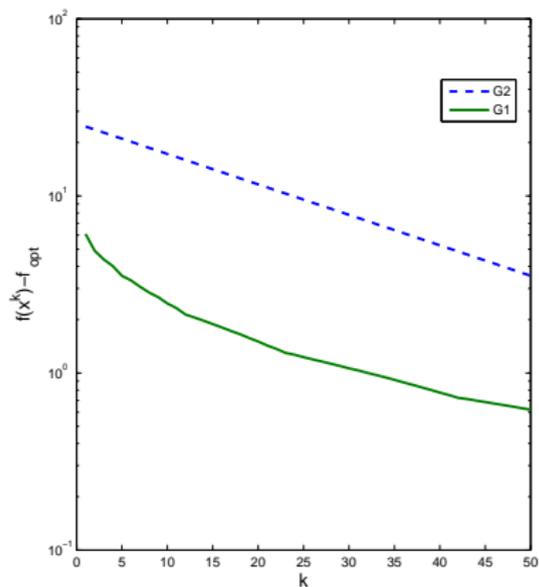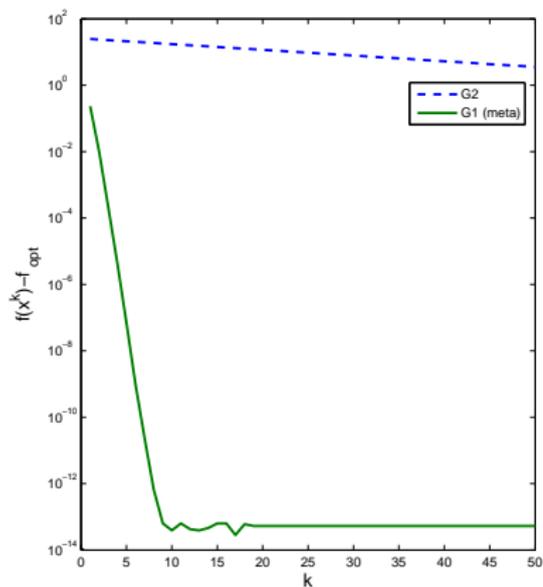
Numerical Example:

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

$n = 10$
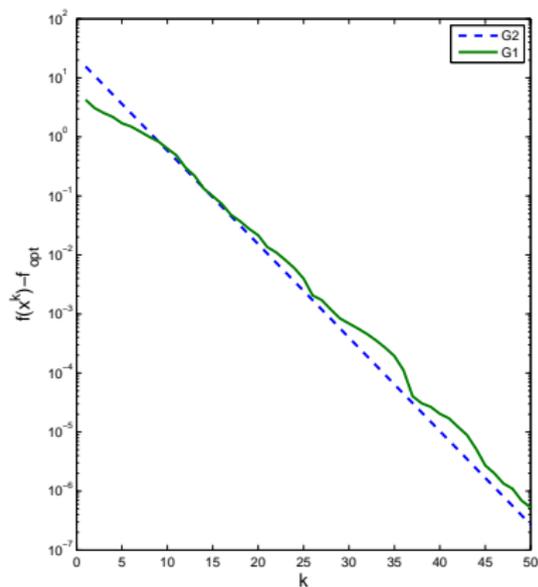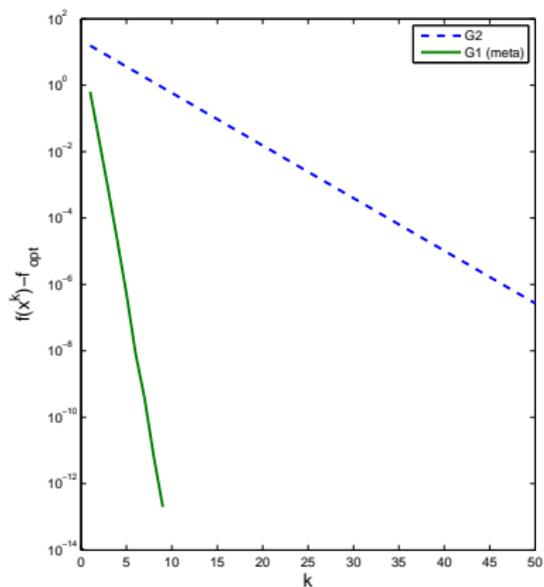
$n = 100$

References

- 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).

Lecture 7 – 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 $\mathbf{x}^{k+1} = \operatorname{prox}_{\frac{1}{L_k} g} \left(\mathbf{y}^k - \frac{1}{L_k} \nabla f(\mathbf{y}^k) \right)$.
>
> (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

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

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

- $\beta L_f \leq L_k \leq \alpha L_f$, 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 $t_k \geq \frac{k+2}{2}$ for all $k \geq 0$.

$O(1/k^2)$ 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 $k \geq 1$,

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

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.

▶ Substituting $\mathbf{x} = t_k^{-1}\mathbf{x}^* + (1 - t_k^{-1})\mathbf{x}^k$ and $\mathbf{y} = \mathbf{y}^k$ and $L = L_k$ in the fundamental prox-grad inequality

$$F(t_k^{-1}\mathbf{x}^* + (1 - t_k^{-1})\mathbf{x}^k) - F(\mathbf{x}^{k+1})$$

$$\geq \frac{L_k}{2}\|\mathbf{x}^{k+1} - (t_k^{-1}\mathbf{x}^* + (1 - t_k^{-1})\mathbf{x}^k)\|^2 - \frac{L_k}{2}\|\mathbf{y}^k - (t_k^{-1}\mathbf{x}^* + (1 - t_k^{-1})\mathbf{x}^k)\|^2$$

$$= \frac{L_k}{2t_k^2}\|t_k\mathbf{x}^{k+1} - (\mathbf{x}^* + (t_k - 1)\mathbf{x}^k)\|^2 - \frac{L_k}{2t_k^2}\|t_k\mathbf{y}^k - (\mathbf{x}^* + (t_k - 1)\mathbf{x}^k)\|^2. \quad (1)$$

Proof Contd.

- By the convexity of F, $F(t_k^{-1}\mathbf{x}^* + (1 - t_k^{-1})\mathbf{x}^k) \leq t_k^{-1}F(\mathbf{x}^*) + (1 - t_k^{-1})F(\mathbf{x}^k)$.

- Using the notation $v_n \equiv F(\mathbf{x}^n) - F_{\mathrm{opt}}$,

$$F(t_k^{-1}\mathbf{x}^* + (1 - t_k^{-1})\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \leq (1 - t_k^{-1})v_k - v_{k+1}. \qquad (2)$$

- By the relation $\mathbf{y}^k = \mathbf{x}^k + \left(\frac{t_{k-1}-1}{t_k}\right)(\mathbf{x}^k - \mathbf{x}^{k-1})$,

$$\|t_k\mathbf{y}^k - (\mathbf{x}^* + (t_k - 1)\mathbf{x}^k)\|^2 = \|t_{k-1}\mathbf{x}^k - (\mathbf{x}^* + (t_{k-1} - 1)\mathbf{x}^{k-1})\|^2 \qquad (3)$$

- Combining (1),(2) and (3),

$$(t_k^2 - t_k)v_k - t_k^2 v_{k+1} \geq \frac{L_k}{2}\|\mathbf{u}^{k+1}\|^2 - \frac{L_k}{2}\|\mathbf{u}^k\|^2,$$

where $\mathbf{u}^n \equiv t_{n-1}\mathbf{x}^n - (\mathbf{x}^* + (t_{n-1} - 1)\mathbf{x}^{n-1})$.

- Since $t_k^2 - t_k = t_{k-1}^2$,

$$\frac{2}{L_k}t_{k-1}^2 v_k - \frac{2}{L_k}t_k^2 v_{k+1} \geq \|\mathbf{u}^{k+1}\|^2 - \|\mathbf{u}^k\|^2.$$

Proof Contd.

- Since $L_k \geq L_{k-1}$,

$$\frac{2}{L_{k-1}} t_{k-1}^2 v_k - \frac{2}{L_k} t_k^2 v_{k+1} \geq \|\mathbf{u}^{k+1}\|^2 - \|\mathbf{u}^k\|^2.$$

- $\|\mathbf{u}^{k+1}\|^2 + \frac{2}{L_k} t_k^2 v_{k+1} \leq \|\mathbf{u}^k\|^2 + \frac{2}{L_{k-1}} t_{k-1}^2 v_k,$
- $\|\mathbf{u}^k\|^2 + \frac{2}{L_{k-1}} t_{k-1}^2 v_k \leq \|\mathbf{u}^1\|^2 + \frac{2}{L_0} t_0^2 v_1 = \|\mathbf{x}^1 - \mathbf{x}^*\|^2 + \frac{2}{L_0}(F(\mathbf{x}^1) - F_{\mathrm{opt}})$
- Substituting $\mathbf{x} = \mathbf{x}^*, \mathbf{y} = \mathbf{y}^0$ and $L = L_0$ in the fundamental prox-grad inequality,

$$\frac{2}{L_0}(F(\mathbf{x}^*) - F(\mathbf{x}^1)) \geq \|\mathbf{x}^1 - \mathbf{x}^*\|^2 - \|\mathbf{y}^0 - \mathbf{x}^*\|^2.$$

- Hence, since $\mathbf{y}^0 = \mathbf{x}^0$,

$$\|\mathbf{x}^1 - \mathbf{x}^*\|^2 + \frac{2}{L_0}(F(\mathbf{x}^1) - F_{\mathrm{opt}}) \leq \|\mathbf{x}^0 - \mathbf{x}^*\|^2.$$

- $\frac{2}{L_{k-1}} t_{k-1}^2 v_k \leq \|\mathbf{u}^k\|^2 + \frac{2}{L_{k-1}} t_{k-1}^2 v_k \leq \|\mathbf{x}^0 - \mathbf{x}^*\|^2.$
- Since $L_{k-1} \leq \alpha L_f$,

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

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 **S**hrinkage/**T**hresholding **A**lgorithm (**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 **I**terative **S**hrinkage/**T**hresholding **A**lgorithm (**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 $\mathbf{x}_{\mathrm{true}} = \mathbf{e}_3 - \mathbf{e}_7$.
- $\mathbf{b} = \mathbf{A}\mathbf{x}_{\mathrm{true}}$.
- ran 200 iterations of ISTA and FISTA with $\mathbf{x}^0 = \mathbf{e}$.

Function Values
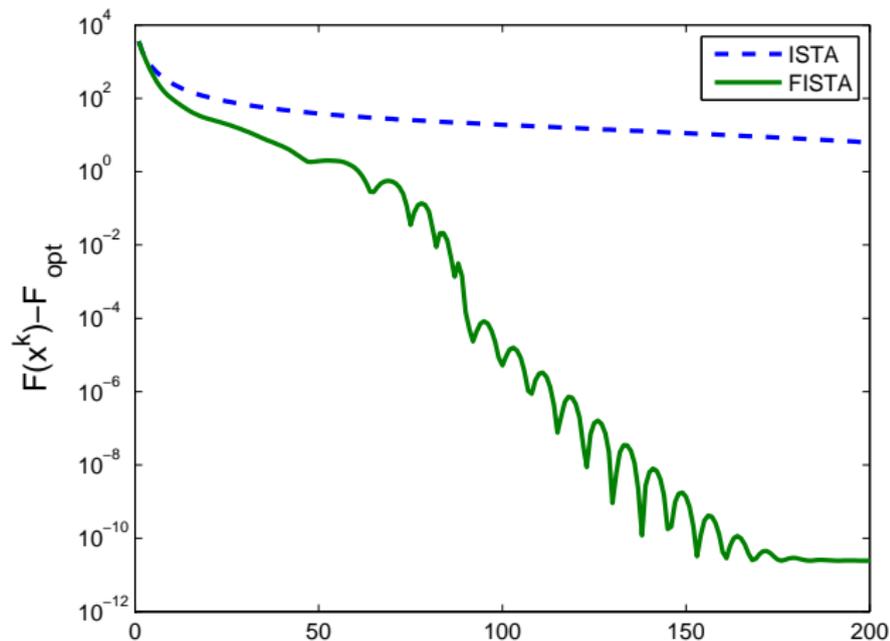
Solutions
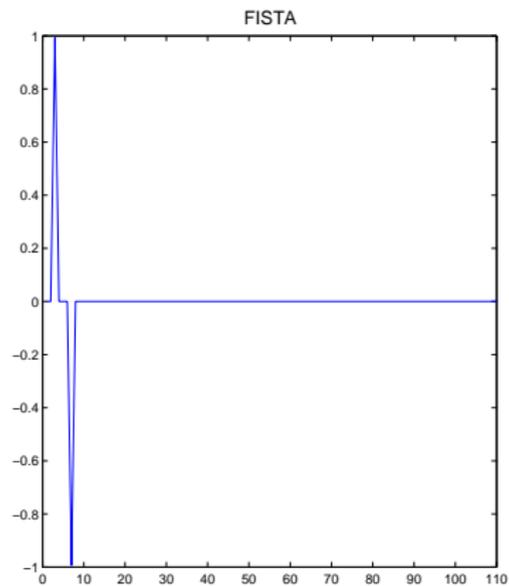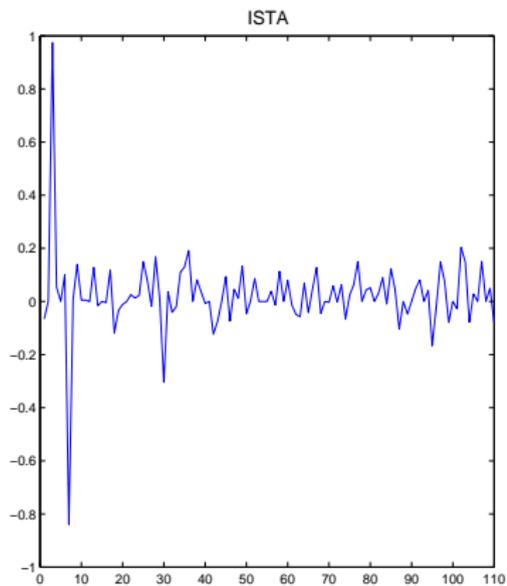
Example 2: Wavelet-Based Image Deblurring

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

- image of size 512x512
- matrix $\mathbf{A}$ 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 $F(\mathbf{x}^k) - F(\mathbf{x}^*)$
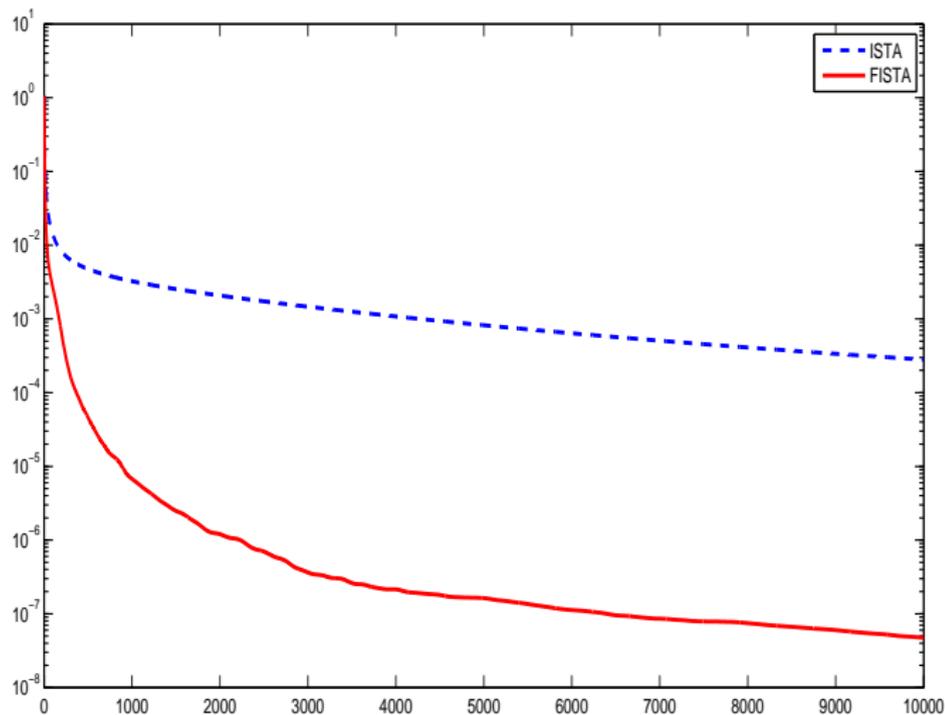
Weighted FISTA

- $\mathbb{E} = \mathbb{R}^n$
- 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):

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

Weighted FISTA

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

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

(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 $\mathbf{Q}$-norm:

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

The convergence result will also be written in term of the $\mathbf{Q}$-norm

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

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

(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:

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

Sketch of Proof of (a)

(b) follows easily from (a). To prove (a),

- $F(z^{n+1}) - F_{\text{opt}} \leq \frac{2L_f \|z^n - x^*\|^2}{(N+1)^2}$.
- $F(z^n) - F_{\text{opt}} \geq \frac{\sigma}{2} \|z^n - x^*\|^2$.
- Hence, $F(z^{n+1}) - F_{\text{opt}} \leq \frac{4\kappa(F(z^n) - F_{\text{opt}})}{(N+1)^2}$.
- Since $N \geq \sqrt{8\kappa} - 1$, it follows that $\frac{4\kappa}{(N+1)^2} \leq \frac{1}{2}$, and hence,

$$F(z^{n+1}) - F_{\text{opt}} \leq \frac{1}{2}(F(z^n) - F_{\text{opt}}).$$

- $F(z^k) - F_{\text{opt}} \leq \left(\frac{1}{2}\right)^k (F(z^0) - F_{\text{opt}})$.
- Need to show that $F(z^0) - F_{\text{opt}} \leq \frac{L_f R^2}{2}$...

References

- 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)

Lecture 8 – 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 $h : \mathbb{E} \to \mathbb{E}$ 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) $h_\mu(\mathbf{x}) \leq h(\mathbf{x}) \leq h_\mu(\mathbf{x}) + \beta\mu$ 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 $q(\mathbf{x}) = \max\limits_{i=1,\ldots,m} \{\mathbf{a}_i^T \mathbf{x} + b_i\}$, 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 $(\|\mathbf{A}\|_{2,2}^2, \log m)$.

Tightness of the Smoothing Parameters

- Let $q(x) = |x|$.
- The function $\sqrt{x^2 + \mu^2} - \mu$ is a $\frac{1}{\mu}$-smooth approximation of q with parameters $(1, 1)$.
- Alternative approach: $q(x) = \max\{x, -x\}$.
- $q_\mu(x) = \mu \log(e^{x/\mu} + e^{-x/\mu}) - \mu \log 2$ is a $\frac{1}{\mu}$-smooth approximation of q with parameters $(2, \log 2)$. Are these the "best" parameters?
- The β-parameter is indeed tight since $\lim_{x \to \infty} q(x) - q_\mu(x) = \mu \log(2)$.
- The α-parameter is not tight since for any x,

$$|q_1''(x)| = \left| \frac{4}{(e^x + e^{-x})^2} \right| \leq 1.$$

  Thus, q_1 is 1-smooth, and hence $q_\mu(\mathbf{x}) = \mu q_1(\mathbf{x}/\mu)$ is $\frac{1}{\mu}$-smooth.
- q_μ is a $\frac{1}{\mu}$-smooth approximation of q with parameters $(1, \log 2)$.

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 $\operatorname{prox}_f = P_C$ 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 $\mathbf{x} \in \mathbb{E}$,

$$\operatorname{prox}_{\mu f}(\mathbf{x}) = \left(1 - \frac{\mu}{\max\{\|\mathbf{x}\|, \mu\}}\right)\mathbf{x}.$$

Therefore,

$$M_f^\mu(\mathbf{x}) = \|\operatorname{prox}_{\mu f}(\mathbf{x})\| + \frac{1}{2\mu}\|\mathbf{x} - \operatorname{prox}_{\mu f}(\mathbf{x})\|^2 = \underbrace{\begin{cases} \frac{1}{2\mu}\|\mathbf{x}\|^2, & \|\mathbf{x}\| \leq \mu, \\ \|\mathbf{x}\| - \frac{\mu}{2}, & \|\mathbf{x}\| > \mu, \end{cases}}_{H_\mu(\mathbf{x})}$$

H_μ - Huber function

Huber Function

H_μ gets smoother as μ increases.
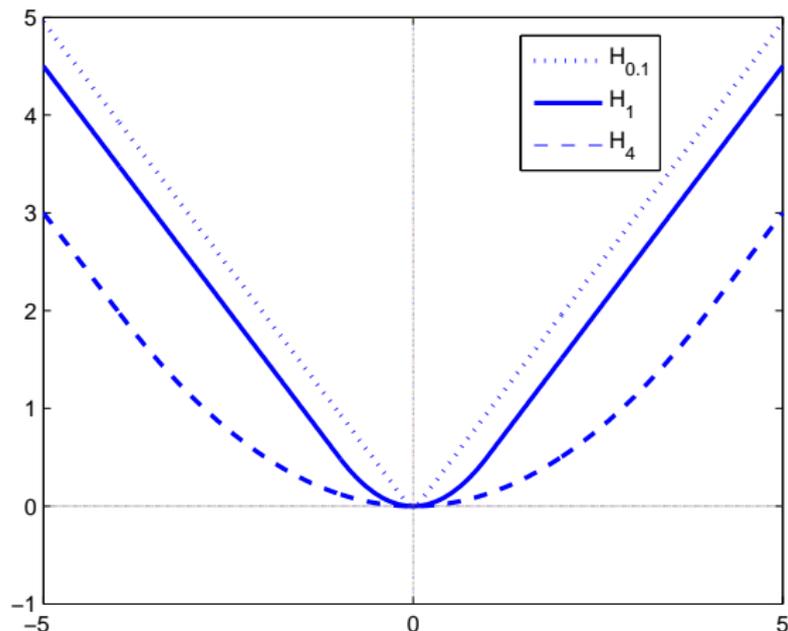
Calculus of Moreau Envelops

- Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper closed and convex function, and let $\lambda, \mu > 0$. Then

$$\lambda M_f^\mu(\mathbf{x}) = M_{\lambda f}^{\mu/\lambda}(\mathbf{x}).$$

- Suppose that $\mathbb{E} = \mathbb{E}_1 \times \mathbb{E}_2 \times \cdots \times \mathbb{E}_m$, and let $f : \mathbb{E} \to (-\infty, \infty]$ be given by

$$f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m) = \sum_{i=1}^m f_i(\mathbf{x}_i), \quad \mathbf{x}_1 \in \mathbb{E}_1, \mathbf{x}_2 \in \mathbb{E}_2, \ldots, \mathbf{x}_m \in \mathbb{E}_m,$$

with $f_i : \mathbb{E}_i \to (-\infty, \infty]$ being a proper closed and convex function for any i. Then for any $\mu > 0$,

$$M_f^\mu(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m) = \sum_{i=1}^m M_{f_i}^\mu(\mathbf{x}_i).$$

Example: Let $f(\mathbf{x}) = \|\mathbf{x}\|_1$. Then $f(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{i=1}^n g(x_i)$, where $g(t) = |t|$. Thus,

$$M_f^\mu(\mathbf{x}) = \sum_{i=1}^n M_g^\mu(x_i) = \sum_{i=1}^n H_\mu(x_i).$$

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

$$\nabla \left(1/2 d_C^2 \right)(\mathbf{x}) = \mathbf{x} - \mathrm{prox}_{\delta_C}(\mathbf{x}) = \mathbf{x} - P_C(\mathbf{x}).$$

- **(smoothability of Huber)** $H_\mu = M_f^\mu$, where $f(\mathbf{x}) = \|\mathbf{x}\|$. Then H_μ is $\frac{1}{\mu}$-smooth and

$$
\begin{aligned}
\nabla H_\mu(\mathbf{x}) &= \frac{1}{\mu} \left(\mathbf{x} - \mathrm{prox}_{\mu f}(\mathbf{x}) \right) = \frac{1}{\mu} \left(\mathbf{x} - \left(1 - \frac{\mu}{\max\{\|\mathbf{x}\|, \mu\}} \right) \mathbf{x} \right) \\
&= \begin{cases} \frac{1}{\mu} \mathbf{x}, & \|\mathbf{x}\| \leq \mu, \\ \frac{\mathbf{x}}{\|\mathbf{x}\|}, & \|\mathbf{x}\| > \mu, \end{cases}
\end{aligned}
$$

Smoothability of Lipschitz Convex Functions

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

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

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

Proof.

- M_h^μ is $\frac{1}{\mu}$-smooth.
- For any $\mathbf{x} \in \mathbb{E}$,

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

- Let $\mathbf{g_x} \in \partial h(\mathbf{x})$. Then $\|\mathbf{g_x}\| \leq \ell_h$, and hence

$$
\begin{aligned}
M_h^\mu(\mathbf{x}) - h(\mathbf{x}) &= \min_{\mathbf{u} \in \mathbb{E}} \left\{ h(\mathbf{u}) - h(\mathbf{x}) + \frac{1}{2\mu} \|\mathbf{u} - \mathbf{x}\|^2 \right\} \\
&\geq \min_{\mathbf{u} \in \mathbb{E}} \left\{ \langle \mathbf{g_x}, \mathbf{u} - \mathbf{x} \rangle + \frac{1}{2\mu} \|\mathbf{u} - \mathbf{x}\|^2 \right\} = -\frac{\mu}{2} \|\mathbf{g_x}\|^2 \geq -\frac{\ell_h^2}{2} \mu,
\end{aligned}
$$

Smoothability of Lipschitz Convex Functions

> 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,

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

  is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(1, \frac{1}{2})$.

- **(smooth approximation of the l_1-norm)** Let $h(\mathbf{x}) = \|\mathbf{x}\|_1$ (over $\mathbb{R}^n$) 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 $h(x) = |x|$

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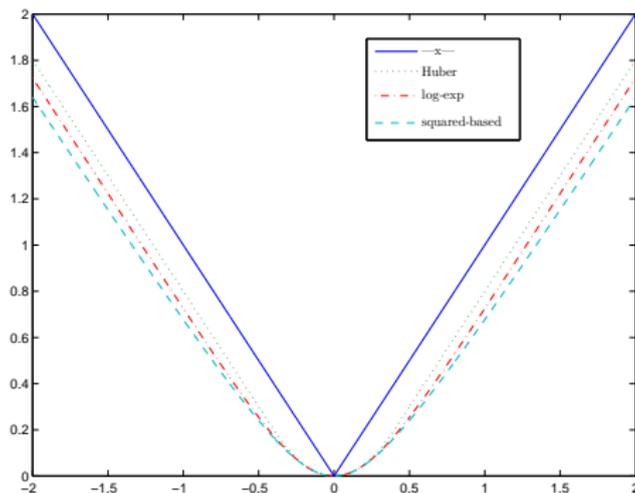
Back to Algorithms - Model and Assumptions

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_{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: $\mathbf{x}^0 \in \text{dom}(g)$, $\mu > 0$.

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

(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 $H(\mathbf{x}^k) - H_{\mathrm{opt}} \leq \varepsilon$.

Proof idea (incomplete and superficial...): $\mu \approx A\varepsilon$, $k \geq \frac{B}{\varepsilon}$. Then

$$H(\mathbf{x}^k) - H_{\mathrm{opt}} \leq H_\mu(\mathbf{x}^k) + \beta\mu - H_{\mu,opt} \leq \frac{C/\mu}{B^2/\varepsilon^2} + \beta\mu \approx \left(\frac{C}{AB^2} + \beta A\right)\varepsilon$$

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 $(\alpha, \beta) = (1, \frac{\ell_h^2}{2})$.
- $\nabla M_h^\mu(\mathbf{x}) = \frac{1}{\mu}(\mathbf{x} - \text{prox}_{\mu h}(\mathbf{x}))$.
- 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

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

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) $\mathbf{x}^{k+1} = P_C(\mathrm{prox}_{\mu h}(\mathbf{y}^k))$;
(b) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
(c) $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{t_k-1}{t_{k+1}}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.

Example

Consider the problem

$$(P_2) \quad \min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \|\mathbf{D}\mathbf{x}\|_1 + \lambda \|\mathbf{x}\|_1 \right\},$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{b} \in \mathbb{R}^m, \mathbf{D} \in \mathbb{R}^{p \times n}$ and $\lambda > 0$.

- Problem (P_2) fits model (P) with $f(\mathbf{x}) = \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, $h(\mathbf{x}) = \|\mathbf{D}\mathbf{x}\|_1$ and $g(\mathbf{x}) = \lambda \|\mathbf{x}\|_1$.
- All assumptions are met: f is convex and L_f-smooth with $L_f = \|\mathbf{A}^T\mathbf{A}\|_{2,2} = \|\mathbf{A}\|_{2,2}^2$, g is proper closed and convex, h is real-valued and convex and the level sets of the objective function are bounded.
- $h(\mathbf{x}) = q(\mathbf{D}\mathbf{x})$, where $q : \mathbb{R}^p \to \mathbb{R}$ is given by $q(\mathbf{y}) = \|\mathbf{y}\|_1$.
- For any $\mu > 0$, $q_\mu(\mathbf{y}) = M_q^\mu(\mathbf{y}) = \sum_{i=1}^p H_\mu(y_i)$ is a $\frac{1}{\mu}$-smooth approximation of q with parameters $(1, \frac{p}{2})$.
- $h_\mu(\mathbf{x}) \equiv q_\mu(\mathbf{D}\mathbf{x})$ is a $\frac{1}{\mu}$-smooth approximation of h with parameters $(\alpha, \beta) = (\|\mathbf{D}\|_{2,2}^2, \frac{p}{2})$.
- To obtain an ε-optimal solution of problem (P), we need to employ the S-FISTA method with $\mu = \sqrt{\frac{\alpha}{\beta}} \frac{\varepsilon}{\sqrt{\alpha\beta} + \sqrt{\alpha\beta + L_f \varepsilon}}$

Example Contd.

- The smooth part is $F_\mu(\mathbf{x}) = f(\mathbf{x}) + M_q^\mu(\mathbf{Dx})$. The nonsmooth part is $g(\mathbf{x}) = \lambda \|\mathbf{x}\|_1$

-
$$
\begin{aligned}
\nabla F_\mu(\mathbf{x}) &= \nabla f(\mathbf{x}) + \mathbf{D}^T \nabla M_q^\mu(\mathbf{Dx}) \\
&= \nabla f(\mathbf{x}) + \tfrac{1}{\mu} \mathbf{D}^T (\mathbf{Dx} - \mathrm{prox}_{\mu q}(\mathbf{Dx})) \\
&= \nabla f(\mathbf{x}) + \tfrac{1}{\mu} \mathbf{D}^T (\mathbf{Dx} - \mathcal{T}_\mu(\mathbf{Dx}))
\end{aligned}
$$

S-FISTA for solving (P)
Initialization: set $\mathbf{y}^0 = \mathbf{x}^0 \in \mathbb{R}^n$, $t_0 = 1$; set μ as in the previous slide and $\tilde{L} = \|\mathbf{A}\|_{2,2}^2 + \frac{\|\mathbf{D}\|_{2,2}^2}{\mu}$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) $\mathbf{x}^{k+1} = \mathcal{T}_{\lambda/\tilde{L}} \left(\mathbf{y}^k - \tfrac{1}{\tilde{L}} (\nabla f(\mathbf{y}^k) + \tfrac{1}{\mu} \mathbf{D}^T (\mathbf{Dy}^k - \mathcal{T}_\mu(\mathbf{Dy}^k))) \right)$;

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

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

References

► 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)

► 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)

Lecture 9 – Block Proximal Gradient Methods

The Model

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

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

Stepsize 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

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

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

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

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

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

References

- 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)

Lecture 10 – Dual-Based Proximal Gradient Methods

Main Model:

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

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 $\hat{\mathbf{x}} \in \mathrm{ri}(\mathrm{dom}(f))$ and $\hat{\mathbf{z}} \in \mathrm{ri}(\mathrm{dom}(g))$ 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 $\mathbb{E} = \mathbb{R}^n$, $\mathbb{V} = \mathbb{R}^p$, $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|^2$,

$$g(\mathbf{z}) = \delta_{\mathrm{Box}[-\infty\mathbf{e},\mathbf{b}]}(\mathbf{z}) = \left\{ \begin{array}{ll} \mathbf{0}, & \mathbf{z} \leq \mathbf{b}, \\ \infty, & \text{else.} \end{array} \right.$$

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,
$\mathbb{E} = \mathbb{R}^n, \mathbb{V} = \mathbb{R}^{n-1}, \mathcal{A}(\mathbf{x}) = \mathbf{Dx}, R(\mathbf{z}) = \lambda\|\mathbf{z}\|_1 (\lambda > 0), \mathbf{D}$ defined by
$\mathbf{Dx} = (x_1 - x_2, x_2 - x_3, \ldots, x_{n-1} - x_n)^T$

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

▶ 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 $\mathbf{x} \mapsto \|\mathbf{Dx}\|_1$ is a one-dimensional total variation function.

▶ Fits model (P) with
$\mathbb{E} = \mathbb{R}^n, \mathbb{V} = \mathbb{R}^{n-1}, f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|^2 (\sigma = 1), g(\mathbf{y}) = \lambda\|\mathbf{y}\|_1, \mathcal{A}(\mathbf{x}) \equiv \mathbf{Dx}$

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. $\mathbf{x}$ and $\mathbf{z}$, we obtain the dual problem

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

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

The dual problem in minimization form:

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

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

Baillon Haddad Theorem

> Theorem [Baillon-Haddad] 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.

Proof of (b)

- Suppose that f is proper closed and σ-strongly convex.
- By the conjugate subgradient theorem
  $\partial f^*(\mathbf{y}) = \underset{\mathbf{x} \in \mathbb{E}}{\operatorname{argmax}}\{\langle \mathbf{x}, \mathbf{y} \rangle - f(\mathbf{x})\}$ for any $\mathbf{y} \in \mathbb{E}$.
- Thus, since f is closed and strongly convex, $\partial f^*(\mathbf{y})$ is a singleton for any $\mathbf{y}$.
- f^* is differentiable over $\mathbb{E}$.
- Take $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{E}$ and denote $\mathbf{v}_1 = \nabla f^*(\mathbf{y}_1), \mathbf{v}_2 = \nabla f^*(\mathbf{y}_2)$.
- By the conjugate subgradient theorem, $\mathbf{y}_1 \in \partial f(\mathbf{v}_1), \mathbf{y}_2 \in \partial f(\mathbf{v}_2)$.
- Thus, $\langle \mathbf{y}_1 - \mathbf{y}_2, \mathbf{v}_1 - \mathbf{v}_2 \rangle \geq \sigma \|\mathbf{v}_1 - \mathbf{v}_2\|^2$,
- $\langle \mathbf{y}_1 - \mathbf{y}_2, \nabla f^*(\mathbf{y}_1) - \nabla f^*(\mathbf{y}_2) \rangle \geq \sigma \|\nabla f^*(\mathbf{y}_1) - \nabla f^*(\mathbf{y}_2)\|^2$,
- By the Cauchy-Schwarz inequality, $\|\nabla f^*(\mathbf{y}_1) - \nabla f^*(\mathbf{y}_2)\| \leq \frac{1}{\sigma} \|\mathbf{y}_1 - \mathbf{y}_2\|_*$.

The Dual Problem

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

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.

Only non-trivial property is the L_F-smoothness of F: by the Baillon-Haddad theorem f^* is $\frac{1}{\sigma}$-smooth. Therefore,

$$\begin{aligned}
\|\nabla F(\mathbf{y}_1) - \nabla F(\mathbf{y}_2)\| &= \|\mathcal{A}(\nabla f^*(\mathcal{A}^T(\mathbf{y}_1))) - \mathcal{A}(\nabla f^*(\mathcal{A}^T(\mathbf{y}_2)))\| \\
&\leq \|\mathcal{A}\| \cdot \|\nabla f^*(\mathcal{A}^T(\mathbf{y}_1)) - \nabla f^*(\mathcal{A}^T(\mathbf{y}_2))\| \\
&\leq \frac{1}{\sigma} \|\mathcal{A}\| \cdot \|\mathcal{A}^T(\mathbf{y}_1) - \mathcal{A}^T(\mathbf{y}_2)\| \\
&\leq \frac{\|\mathcal{A}\| \cdot \|\mathcal{A}^T\|}{\sigma} \|\mathbf{y}_1 - \mathbf{y}_2\| = \frac{\|\mathcal{A}\|^2}{\sigma} \|\mathbf{y}_1 - \mathbf{y}_2\|,
\end{aligned}$$

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** $(k \geq 0)$:
>
> $$\mathbf{y}^{k+1} = \mathrm{prox}_{\frac{1}{L}G}\left(\mathbf{y}^k - \frac{1}{L}\nabla F(\mathbf{y}^k)\right).$$

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 $\mathbf{y}^*$ $k \geq 1$,

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

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

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

holds if and only if

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

where

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

Proof.

▶ By the conjugate subgradient theorem,

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

▶ Therefore, since $\nabla F(\mathbf{v}) = \mathcal{A}(\nabla f^*(\mathcal{A}^T(\mathbf{v}) + \mathbf{b})) = \mathcal{A}(\tilde{\mathbf{x}})$

$$\mathbf{y} = \mathrm{prox}_{\frac{1}{L} G} \left(\mathbf{v} - \frac{1}{L} \mathcal{A}(\tilde{\mathbf{x}}) \right). \qquad (2)$$

▶ For any $\mathbf{z} \in \mathbb{V}$, $\mathrm{prox}_{\frac{1}{L} G}(\mathbf{z}) = -\mathrm{prox}_{\frac{1}{L} g^*}(-\mathbf{z})$.

▶ Using the the extended Moreau decomposition formula, we finally obtain that

$$
\begin{aligned}
\mathbf{y} &= \mathrm{prox}_{\frac{1}{L} G} \left(\mathbf{v} - \frac{1}{L} \mathcal{A}(\tilde{\mathbf{x}}) \right) = -\mathrm{prox}_{\frac{1}{L} g^*} \left(\frac{1}{L} \mathcal{A}(\tilde{\mathbf{x}}) - \mathbf{v} \right) \\
&= -\left[\frac{1}{L} \mathcal{A}(\tilde{\mathbf{x}}) - \mathbf{v} - \frac{1}{L} \mathrm{prox}_{Lg}(\mathcal{A}(\tilde{\mathbf{x}}) - L\mathbf{v}) \right] \\
&= \mathbf{v} - \frac{1}{L} \mathcal{A}(\tilde{\mathbf{x}}) + \frac{1}{L} \mathrm{prox}_{Lg}(\mathcal{A}(\tilde{\mathbf{x}}) - L\mathbf{v}).
\end{aligned}
$$

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

▶ 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

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

Proof.

- ▶ The primal is the same as $\min_{\mathbf{x} \in \mathbb{E}, \mathbf{z} \in \mathbb{V}} \{ f(\mathbf{x}) + g(\mathbf{z}) : \mathcal{A}(\mathbf{x}) - \mathbf{z} = \mathbf{0} \}$.
- ▶ Lagrangian: $L(\mathbf{x}, \mathbf{z}; \mathbf{y}) = f(\mathbf{x}) - \langle \mathcal{A}^T(\mathbf{y}), \mathbf{x} \rangle + g(\mathbf{z}) + \langle \mathbf{y}, \mathbf{z} \rangle$.
- ▶ In particular, $L(\mathbf{x}, \mathbf{z}; \bar{\mathbf{y}}) = h(\mathbf{x}) + s(\mathbf{z})$, where

$$\begin{aligned} h(\mathbf{x}) &= f(\mathbf{x}) - \langle \mathcal{A}^T(\bar{\mathbf{y}}), \mathbf{x} \rangle, \\ s(\mathbf{z}) &= g(\mathbf{z}) + \langle \bar{\mathbf{y}}, \mathbf{z} \rangle. \end{aligned}$$

Proof Contd.

- h is σ-strongly convex and $\bar{\mathbf{x}} \in \operatorname{argmin} h(\mathbf{x})$. Thus, $h(\mathbf{x}) - h(\bar{\mathbf{x}}) \geq \frac{\sigma}{2}\|\mathbf{x} - \bar{\mathbf{x}}\|^2$.

- The relation $\bar{\mathbf{y}} \in \operatorname{dom}(G)$ is the same as $-\bar{\mathbf{y}} \in \operatorname{dom}(g^*)$. Therefore,

$$\min_{\mathbf{z} \in \mathbb{V}}\{g(\mathbf{z}) + \langle \bar{\mathbf{y}}, \mathbf{z} \rangle\} = \min_{\mathbf{z} \in \mathbb{V}} s(\mathbf{z}) > -\infty.$$

- Let $\varepsilon > 0$. Then there exists $\bar{\mathbf{z}}_\varepsilon$ for which $s(\bar{\mathbf{z}}_\varepsilon) \leq \min_{\mathbf{z} \in \mathbb{V}} s(\mathbf{z}) + \varepsilon$.

- For all $\mathbf{x} \in \operatorname{dom}(f)$ and $\mathbf{z} \in \operatorname{dom}(g)$,

$$L(\mathbf{x}, \mathbf{z}; \bar{\mathbf{y}}) - L(\bar{\mathbf{x}}, \bar{\mathbf{z}}_\varepsilon; \bar{\mathbf{y}}) = h(\mathbf{x}) - h(\bar{\mathbf{x}}) + s(\mathbf{z}) - s(\bar{\mathbf{z}}_\varepsilon) \geq \frac{\sigma}{2}\|\mathbf{x} - \bar{\mathbf{x}}\|^2 - \varepsilon.$$

- Substituting $\mathbf{x} = \mathbf{x}^*, \mathbf{z} = \mathbf{z}^* \equiv \mathcal{A}(\mathbf{x}^*)$, then
  $L(\mathbf{x}^*, \mathbf{z}^*; \bar{\mathbf{y}}) = f(\mathbf{x}^*) + g(\mathcal{A}(\mathbf{x}^*)) = f_{\mathrm{opt}} = q_{\mathrm{opt}}$

- $q_{\mathrm{opt}} - L(\bar{\mathbf{x}}, \bar{\mathbf{z}}_\varepsilon; \bar{\mathbf{y}}) \geq \frac{\sigma}{2}\|\mathbf{x}^* - \bar{\mathbf{x}}\|^2 - \varepsilon.$

- By the definition of the dual objective function,
  $L(\bar{\mathbf{x}}, \bar{\mathbf{z}}_\varepsilon; \bar{\mathbf{y}}) \geq \min_{\mathbf{x} \in \mathbb{E}, \mathbf{z} \in \mathbb{V}} L(\mathbf{x}, \mathbf{z}; \bar{\mathbf{y}}) = q(\bar{\mathbf{y}}),$

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

- Since the above inequality holds for any $\varepsilon > 0$, the result follows.

$O(1/k)$ 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) $\mathbf{y}^{k+1} = \text{prox}_{\frac{1}{L} G} \left(\mathbf{w}^k - \frac{1}{L} \nabla F(\mathbf{w}^k) \right);$
>   (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 $k \geq 1$,

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

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

(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.

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

Example 1: Orthogonal Projection onto a Polyhedral set

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

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

$$g(\mathbf{z}) = \delta_{\mathrm{Box}[-\infty\mathbf{e}, \mathbf{b}]}(\mathbf{z}) = \left\{ \begin{array}{ll} \mathbf{0}, & \mathbf{z} \leq \mathbf{b}, \\ \infty, & \text{else.} \end{array} \right.$$

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 $\mathbf{y} \in \mathbb{R}^p$;

▶ $\operatorname{prox}_{Lg}(\mathbf{z}) = P_{\mathrm{Box}[-\infty\mathbf{e},\mathbf{b}]}(\mathbf{z}) = \min\{\mathbf{z}, \mathbf{b}\}$.

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
  $\mathcal{A} : \mathbb{E} \to \mathbb{V}, \mathcal{A}(\mathbf{z}) = \underbrace{(\mathbf{z}, \mathbf{z}, \ldots, \mathbf{z})}_{p \text{ times}}$
- $\operatorname*{argmax}_{\mathbf{x}}\{\langle \mathbf{v}, \mathbf{x} \rangle - f(\mathbf{x})\} = \mathbf{v} + \mathbf{d}$ 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.
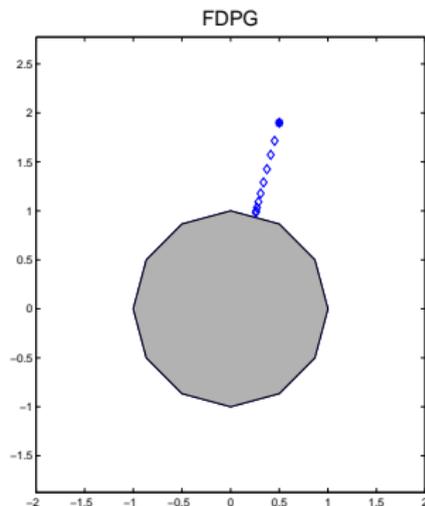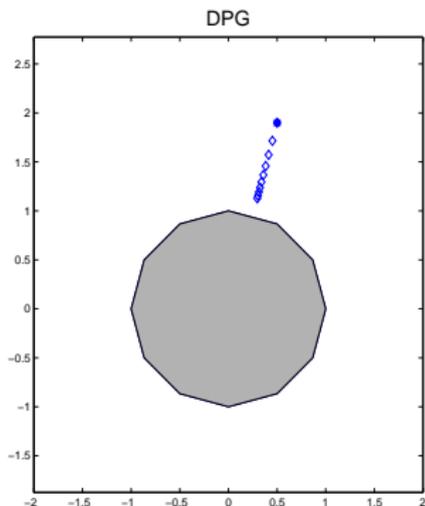
Example 2: One-Dimensional Total Variation Denoising

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

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

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

- $n = 1000$
- **d** is a noisy measurement of a step function.
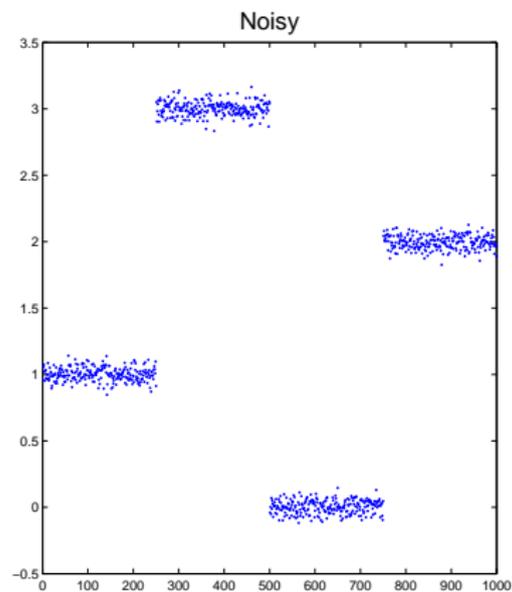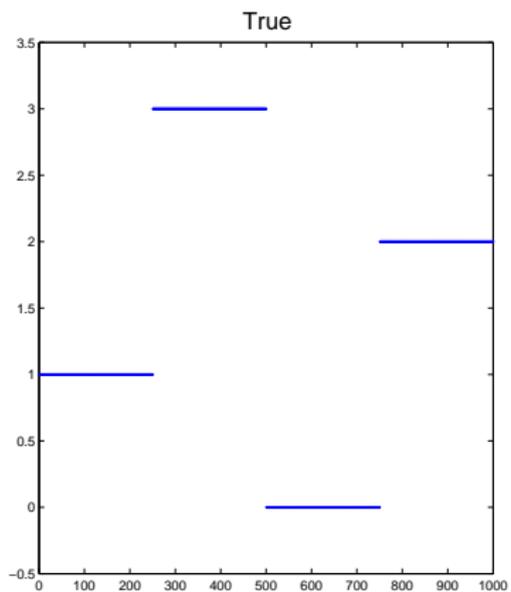
Numerical Example Contd.

▶ 100 iterations of Algorithms 6 (DPG) and 7 (FDPG) initialized with $\mathbf{y}^0 = \mathbf{0}$.
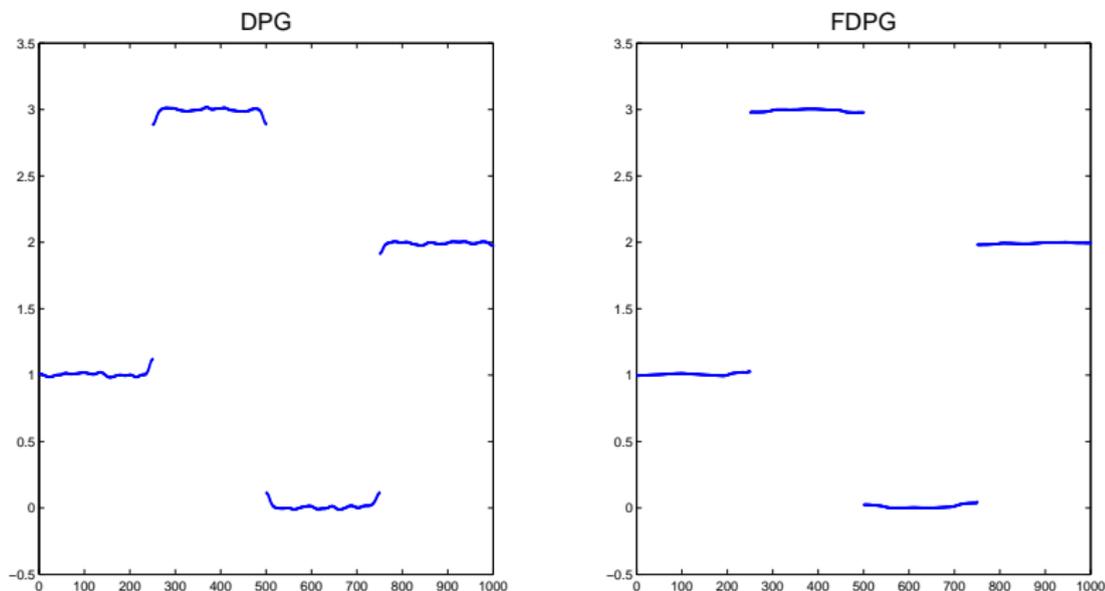
▶ Objective function values of the DPG and FDPG methods after 100 iterations are 9.1667 and 8.4621 respectively; the optimal value is 8.3031.

The Dual Block Proximal Gradient Method

The Model

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

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

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

  (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):

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

- 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:

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

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 $\mathbf{y}_j^{k+1} = \begin{cases} \text{prox}_{\sigma G_{i_k}} \left(\mathbf{y}_{i_k}^k - \sigma \nabla f^* (\sum_{j=1}^p \mathbf{y}_j^k) \right), & j = i_k, \\ \mathbf{y}_j^k, & j \neq i_k. \end{cases}$

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

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

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

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

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

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

Randomized:

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

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

Accelerated DBPG?

- The deterministic and randomized DBPG methods share an $O(1/k)$ rate of convergence.
- Is it possible to accelerate them into an $O(1/k^2)$ method?
- The answer is *yes* in the 2-block case ($p = 2$).

The Model.
$$f_{\text{opt}} = \min_{\mathbf{x} \in \mathbb{E}} \{F(\mathbf{x}) \equiv f(\mathbf{x}) + g_1(\mathbf{x}) + g_2(\mathbf{x})\}.$$

We can rewrite the problem as

$$\min_{\mathbf{x} \in \mathbb{E}} \{\tilde{f}(\mathbf{x}) + g_2(\mathbf{x})\},$$

where $\tilde{f} = f + g_1$.is proper closed and σ-strongly convex and g_2 is proper closed and convex.

- The accelerated dual block proximal gradient (ADBPG) is the FDPG method with stepsize σ employed on (P) with $f = \tilde{f}$, $g = g_2$, $\mathcal{A} = \mathcal{I}$.

The ADBPG Method

ADBPG Method

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

(a) $\mathbf{u}^k = \underset{\mathbf{u}}{\operatorname{argmax}} \left\{ \langle \mathbf{u}, \mathbf{w}^k \rangle - f(\mathbf{u}) - g_1(\mathbf{u}) \right\}$;

(b) $\mathbf{y}^{k+1} = \mathbf{w}^k - \sigma \mathbf{u}^k + \sigma \operatorname{prox}_{g_2/\sigma}(\mathbf{u}^k - \mathbf{w}^k/\sigma)$;

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

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

► $\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{4\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{\sigma^2(k+1)^2}$, where $\mathbf{x}^k = \underset{\mathbf{x}}{\operatorname{argmax}} \left\{ \langle \mathbf{x}, \mathbf{y}^k \rangle - f(\mathbf{x}) - g_1(\mathbf{x}) \right\}$.

► If $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|^2$, then step (a) amounts to $\mathbf{u}^k = \operatorname{prox}_{g_1}(\mathbf{d} + \mathbf{w}^k)$.

One-Dimensional Total Variation Denoising

Problem.

$$(T) \quad f_{\mathrm{opt}} = \min_{\mathbf{x} \in \mathbb{R}^n} \left\{ F(\mathbf{x}) \equiv \frac{1}{2} \|\mathbf{x} - \mathbf{d}\|_2^2 + \lambda \sum_{i=1}^{n-1} |x_{i-1} - x_i| \right\},$$

where $\mathbf{d} \in \mathbb{R}^n$ and $\lambda > 0$.

▶ (T) can be written as $\min_{\mathbf{x} \in \mathbb{R}^n} \{ f(\mathbf{x}) + g_1(\mathbf{x}) + g_2(\mathbf{x}) \}$, where

$$
\begin{aligned}
f(\mathbf{x}) &= \frac{1}{2} \|\mathbf{x} - \mathbf{d}\|_2^2, \\
g_1(\mathbf{x}) &= \lambda \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} |x_{2i-1} - x_{2i}|, \\
g_2(\mathbf{x}) &= \lambda \sum_{i=1}^{\lfloor \frac{n-1}{2} \rfloor} |x_{2i} - x_{2i+1}|.
\end{aligned}
$$

Computing $\operatorname{prox}_{g_1}$ and $\operatorname{prox}_{g_2}$

▶ The prox of the two-dimensional function $h(y, z) = |y - z|$ is given by

$$
\begin{aligned}
\operatorname{prox}_{\lambda h}(y, z) &= (y, z) + \frac{1}{2\lambda^2}(\mathcal{T}_{2\lambda^2}(\lambda y - \lambda z) - \lambda y + \lambda z)(\lambda, -\lambda) \\
&= (y, z) + \frac{1}{2}([|y - z| - 2\lambda]_+ \operatorname{sgn}(y - z) - y + z)(1, -1).
\end{aligned}
$$

▶ Using the separability of g_1 w.r.t. $\{x_1, x_2\}, \{x_3, x_4\}, \ldots,$

$$\operatorname{prox}_{g_1}(\mathbf{x}) = \mathbf{x} + \frac{1}{2}\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor}([|x_{2i-1} - x_{2i}| - 2\lambda]_+ \operatorname{sgn}(x_{2i-1} - x_{2i}) - x_{2i-1} + x_{2i})(\mathbf{e}_{2i-1} - \mathbf{e}_{2i}),$$

▶ Similarly,

$$\operatorname{prox}_{g_2}(\mathbf{x}) = \mathbf{x} + \frac{1}{2}\sum_{i=1}^{\lfloor \frac{n-1}{2} \rfloor}([|x_{2i} - x_{2i+1}| - 2\lambda]_+ \operatorname{sgn}(x_{2i} - x_{2i+1}) - x_{2i} + x_{2i+1})(\mathbf{e}_{2i} - \mathbf{e}_{2i+1}).$$

▶ The ADBPG method and Algorithm 9 can be employed on problem (T). The computational effort per iteration in both methods is almost identical.

Numerical Example

- ▶ Run of 1000 iterations of both algorithms.
- ▶ 1D TV denoising problem.
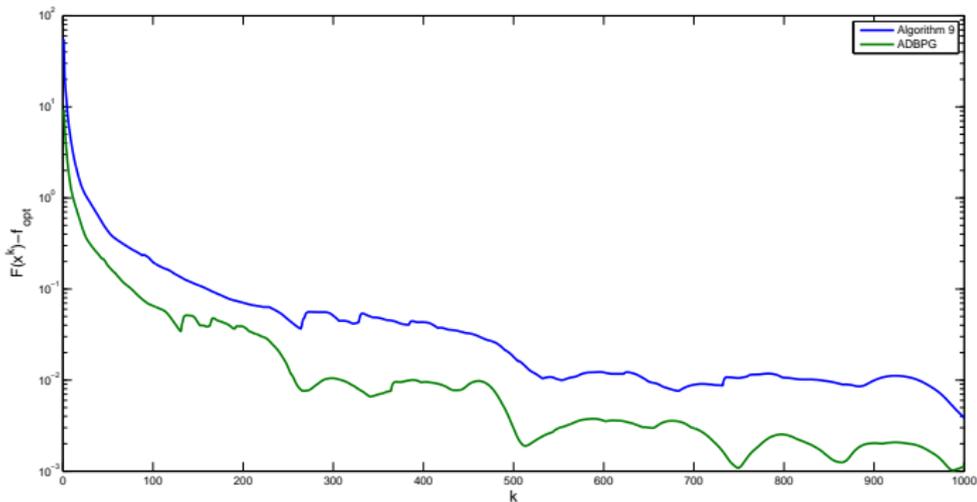
References

▶ 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)

Appendix B

Tables

Support Functions

C	$\sigma_C(\mathbf{y})$	assumptions	reference
$\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$	$\max_{i=1,2,\ldots,n}\langle\mathbf{b}_i,\mathbf{y}\rangle$	$\mathbf{b}_i \in \mathbb{E}$	Example 2.25
K	$\delta_{K^\circ}(\mathbf{y})$	K – cone	Example 2.26
$\mathbb{R}^n_+$	$\delta_{\mathbb{R}^n_-}(\mathbf{y})$	$\mathbb{E} = \mathbb{R}^n$	Example 2.27
Δ_n	$\max\{y_1, y_2, \ldots, y_n\}$	$\mathbb{E} = \mathbb{R}^n$	Example 2.36
$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{0}\}$	$\delta_{\{\mathbf{A}^T\boldsymbol{\lambda}:\boldsymbol{\lambda}\in\mathbb{R}^m_+\}}(\mathbf{y})$	$\mathbb{E} = \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{m\times n}$	Example 2.29
$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{B}\mathbf{x} = \mathbf{b}\}$	$\langle\mathbf{y},\mathbf{x}_0\rangle + \delta_{\text{Range}(\mathbf{B}^T)}(\mathbf{y})$	$\mathbb{E} = \mathbb{R}^n$, $\mathbf{B} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{B}\mathbf{x}_0 = \mathbf{b}$	Example 2.30
$B_{\|\cdot\|}[\mathbf{0},1]$	$\|\mathbf{y}\|_*$	-	Example 2.31

Weak Subdifferential Results

function	weak result	setting	reference
$-q$ - minus dual function	$-\mathbf{g}(\mathbf{x}_0) \in \partial(-q)(\boldsymbol{\lambda}_0)$	$q(\boldsymbol{\lambda}) = \min_{\mathbf{x}\in X} f(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{g}(\mathbf{x})$, $f : \mathbb{E} \to \mathbb{R}$, $\mathbf{g} : \mathbb{E} \to \mathbb{R}^m$, $\mathbf{x}_0$ - a minimizer of $f(\mathbf{x}) + \boldsymbol{\lambda}_0^T\mathbf{g}(\mathbf{x})$ over X	Example 3.7
$f(\mathbf{X}) = \lambda_{\max}(\mathbf{X})$	$\mathbf{v}\mathbf{v}^T \in \partial f(\mathbf{X})$	$f : \mathbb{S}^n \to \mathbb{R}$, $\mathbf{v}$ – normalized maximum eigenvector of $X \in \mathbb{S}^n$	Example 3.8
$f(\mathbf{x}) = \|\mathbf{x}\|_1$	$\text{sgn}(\mathbf{x}) \in \partial f(\mathbf{x})$	$\mathbb{E} = \mathbb{R}^n$	Example 3.42
$f(\mathbf{x}) = \lambda_{\max}(\mathbf{A}_0 + \sum_{i=1}^m x_i\mathbf{A}_i)$	$(\tilde{\mathbf{y}}^T\mathbf{A}_i\tilde{\mathbf{y}})_{i=1}^m \in \partial f(\mathbf{x})$	$\tilde{\mathbf{y}}$ - normalized maximum eigenvector of $\mathbf{A}_0 + \sum_{i=1}^m x_i\mathbf{A}_i$	Example 3.56

Strong Subdifferential Results

$f(\mathbf{x})$	$\partial f(\mathbf{x})$	assumptions	reference
$\|\mathbf{x}\|$	$B_{\|\cdot\|_*}[\mathbf{0},1]$	$\mathbf{x}=\mathbf{0}$	Example 3.3
$\|\mathbf{x}\|_1$	$\left\{\sum_{i\in I_{\neq}(\mathbf{x})}\mathrm{sgn}(x_i)\mathbf{e}_i+\sum_{i\in I_0(\mathbf{x})}[-\mathbf{e}_i,\mathbf{e}_i]\right\}$	$\mathbb{E}=\mathbb{R}^n$, $I_{\neq}(\mathbf{x})=\{i:x_i\neq 0\}$, $I_0(\mathbf{x})=\{i:x_i=0\}$.	Example 3.41
$\|\mathbf{x}\|_2$	$\begin{cases}\left\{\frac{\mathbf{x}}{\|\mathbf{x}\|_2}\right\}, & \mathbf{x}\neq\mathbf{0},\\ B_{\|\cdot\|_2}[\mathbf{0},1] & \mathbf{x}=\mathbf{0}.\end{cases}$	$\mathbb{E}=\mathbb{R}^n$	Example 3.34
$\|\mathbf{x}\|_\infty$	$\left\{\sum_{i\in I(\mathbf{x})}\lambda_i\mathrm{sgn}(x_i)\mathbf{e}_i:\begin{array}{c}\sum_{i\in I(\mathbf{x})}\lambda_i=1\\\lambda_i\geq 0\end{array}\right\}$	$\mathbb{E}=\mathbb{R}^n$, $I(\mathbf{x})=\{i:\|\mathbf{x}\|_\infty=|x_i|\}$, $\mathbf{x}\neq\mathbf{0}$	Example 3.52
$\max(\mathbf{x})$	$\left\{\sum_{i\in I(\mathbf{x})}\lambda_i\mathbf{e}_i:\sum_{i\in I(\mathbf{x})}\lambda_i=1,\lambda_i\geq 0\right\}$	$\mathbb{E}=\mathbb{R}^n$, $I(\mathbf{x})=\{i:\max(\mathbf{x})=x_i\}$	Example 3.51
$\max(\mathbf{x})$	Δ_n	$\mathbb{E}=\mathbb{R}^n$, $\mathbf{x}=\alpha\mathbf{e}$ for some $\alpha\in\mathbb{R}$	Example 3.51
$\delta_S(\mathbf{x})$	$N_S(\mathbf{x})$	$S\subseteq\mathbb{E}$	Example 3.5
$\delta_{B[\mathbf{0},1]}(\mathbf{x})$	$\begin{cases}\{\mathbf{y}\in\mathbb{E}^*:\|\mathbf{y}\|_*\leq\langle\mathbf{y},\mathbf{x}\rangle\}, & \|\mathbf{x}\|\leq 1,\\ \emptyset, & \|\mathbf{x}\|>1.\end{cases}$		Example 3.6
$\|\mathbf{A}\mathbf{x}+\mathbf{b}\|_1$	$\sum_{i\in I_{\neq}(\mathbf{x})}\mathrm{sgn}(\mathbf{a}_i^T\mathbf{x}+b_i)\mathbf{a}_i+\sum_{i\in I_0(\mathbf{x})}[-\mathbf{a}_i,\mathbf{a}_i]$	$\mathbb{E}=\mathbb{R}^n$, $\mathbf{A}\in\mathbb{R}^{m\times n}$, $\mathbf{b}\in\mathbb{R}^m$, $I_{\neq}(\mathbf{x})=\{i:\mathbf{a}_i^T\mathbf{x}+b_i\neq 0\}$, $I_0(\mathbf{x})=\{i:\mathbf{a}_i^T\mathbf{x}+b_i=0\}$	Example 3.44
$\|\mathbf{A}\mathbf{x}+\mathbf{b}\|_2$	$\begin{cases}\frac{\mathbf{A}^T(\mathbf{A}\mathbf{x}+\mathbf{b})}{\|\mathbf{A}\mathbf{x}+\mathbf{b}\|_2}, & \mathbf{A}\mathbf{x}+\mathbf{b}\neq\mathbf{0},\\ \mathbf{A}^T B_{\|\cdot\|_2}[\mathbf{0},1], & \mathbf{A}\mathbf{x}+\mathbf{b}=\mathbf{0}.\end{cases}$	$\mathbb{E}=\mathbb{R}^n$, $\mathbf{A}\in\mathbb{R}^{m\times n}$, $\mathbf{b}\in\mathbb{R}^m$	Example 3.45
$\|\mathbf{A}\mathbf{x}+\mathbf{b}\|_\infty$	$\left\{\sum_{i\in I_{\mathbf{x}}}\lambda_i\mathrm{sgn}(\mathbf{a}_i^T\mathbf{x}+b_i)\mathbf{a}_i:\begin{array}{c}\sum_{i\in I_{\mathbf{x}}}\lambda_i=1\\\lambda_i\geq 0\end{array}\right\}$	$\mathbb{E}=\mathbb{R}^n$, $\mathbf{A}\in\mathbb{R}^{m\times n}$, $\mathbf{b}\in\mathbb{R}^m$, $I_{\mathbf{x}}=\{i:\|\mathbf{A}\mathbf{x}+\mathbf{b}\|_\infty=|\mathbf{a}_i^T\mathbf{x}+b_i|\}$, $\mathbf{A}\mathbf{x}+\mathbf{b}\neq\mathbf{0}$	Example 3.54
$\|\mathbf{A}\mathbf{x}+\mathbf{b}\|_\infty$	$\mathbf{A}^T B_{\|\cdot\|_1}[\mathbf{0},1]$	same as above but with $\mathbf{A}\mathbf{x}+\mathbf{b}=\mathbf{0}$	Example 3.54
$\max_i\{\mathbf{a}_i^T\mathbf{x}+\mathbf{b}\}$	$\left\{\sum_{i\in I(\mathbf{x})}\lambda_i\mathbf{a}_i:\sum_{i\in I(\mathbf{x})}\lambda_i=1,\lambda_i\geq 0\right\}$	$\mathbb{E}=\mathbb{R}^n$, $\mathbf{a}_i\in\mathbb{R}^n$, $b_i\in\mathbb{R}$, $I(\mathbf{x})=\{i:f(\mathbf{x})=\mathbf{a}_i^T\mathbf{x}+b_i\}$	Example 3.53
$\frac{1}{2}d_C(\mathbf{x})^2$	$\{\mathbf{x}-P_C(\mathbf{x})\}$	$\mathbb{E}$ - Euclidean	Example 3.31
$d_C(\mathbf{x})$	$\begin{cases}\left\{\frac{\mathbf{x}-P_C(\mathbf{x})}{d_C(\mathbf{x})}\right\}, & \mathbf{x}\notin C,\\ N_C(\mathbf{x}),\cap B[\mathbf{0},1] & \mathbf{x}\in C.\end{cases}$	$\mathbb{E}$ - Euclidean	Example 3.49

Conjugate Calculus Rules

$g(\mathbf{x})$	$g^*(\mathbf{y})$	reference
$\sum_{i=1}^m f_i(\mathbf{x}_i)$	$\sum_{i=1}^m f_i^*(\mathbf{y}_i)$	Theorem 4.12
$\alpha f(\mathbf{x})\ \ (\alpha > 0)$	$\alpha f^*(\mathbf{y}/\alpha)$	Theorem 4.14
$\alpha f(\mathbf{x}/\alpha)\ \ (\alpha > 0)$	$\alpha f^*(\mathbf{y})$	Theorem 4.14
$f(\mathcal{A}(\mathbf{x}-\mathbf{a})) + \langle \mathbf{b}, \mathbf{x}\rangle + c$	$f^*\left((\mathcal{A}^T)^{-1}(\mathbf{y}-\mathbf{b})\right) + \langle \mathbf{a}, \mathbf{y}\rangle - c - \langle \mathbf{a}, \mathbf{b}\rangle$	Theorem 4.13

Conjugate Functions

f	$\mathrm{dom}(f)$	f^*	assumptions	reference
e^x	$\mathbb{R}$	$y\log y - y\ (\mathrm{dom}(f^*) = \mathbb{R}_+)$	–	Section 4.4.1
$-\log x$	$\mathbb{R}_{++}$	$-1 - \log(-y)\ (\mathrm{dom}(f^*) = \mathbb{R}_{--})$	–	Section 4.4.2
$\max\{1-x, 0\}$	$\mathbb{R}$	$y + \delta_{[-1,0]}(y)$	–	Section 4.4.3
$\frac{1}{p}|x|^p$	$\mathbb{R}$	$\frac{1}{q}|y|^q$	$p > 1, \frac{1}{p} + \frac{1}{q} = 1$	Section 4.4.4
$-\frac{x^p}{p}$	$\mathbb{R}_+$	$-\frac{(-y)^q}{q}\ (\mathrm{dom}(f^*) = \mathbb{R}_{--})$	$0 < p < 1,\ \frac{1}{p} + \frac{1}{q} = 1$	Section 4.4.5
$\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$	$\mathbb{R}^n$	$\frac{1}{2}(\mathbf{y}-\mathbf{b})^T\mathbf{A}^{-1}(\mathbf{y}-\mathbf{b}) - c$	$\mathbf{A} \in \mathbb{S}_{++}^n,\ \mathbf{b} \in \mathbb{R}^n, c \in \mathbb{R}$	Section 4.4.6
$\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$	$\mathbb{R}^n$	$\frac{1}{2}(\mathbf{y}-\mathbf{b})^T\mathbf{A}^\dagger(\mathbf{y}-\mathbf{b}) - c$ $(\mathrm{dom}(f^*) = \mathbf{b} + \mathrm{Range}(\mathbf{A}))$	$\mathbf{A} \in \mathbb{S}_+^n,\ \mathbf{b} \in \mathbb{R}^n, c \in \mathbb{R}$	Section 4.4.7
$\sum_{i=1}^n x_i \log x_i$	$\mathbb{R}_+^n$	$\sum_{i=1}^n e^{y_i - 1}$	–	Section 4.4.8
$\sum_{i=1}^n x_i \log x_i$	Δ_n	$\log\left(\sum_{i=1}^n e^{y_i}\right)$	–	Section 4.4.10
$-\sum_{i=1}^n \log x_i$	$\mathbb{R}_{++}^n$	$-n - \sum_{i=1}^n \log(-y_i)$ $(\mathrm{dom}(f^*) = \mathbb{R}_{--}^n)$	–	Section 4.4.9
$\log\left(\sum_{i=1}^n e^{x_i}\right)$	$\mathbb{R}^n$	$\sum_{i=1}^n y_i \log y_i$ $(\mathrm{dom}(f^*) = \Delta_n)$	–	Section 4.4.11
$\max_i\{x_i\}$	$\mathbb{R}^n$	$\delta_{\Delta_n}(\mathbf{y})$	–	Example 4.10
$\delta_C(\mathbf{x})$	C	$\sigma_C(\mathbf{y})$	$C \subseteq \mathbb{E}$	Example 4.2
$\sigma_C(\mathbf{x})$	$\mathrm{dom}(\sigma_C)$	$\delta_{\mathrm{cl}(\mathrm{conv}(C))}(\mathbf{y})$	$C \subseteq \mathbb{E}$	Example 4.9
$\|\mathbf{x}\|$	$\mathbb{E}$	$\delta_{B_{\|\cdot\|_*}[\mathbf{0},1]}(\mathbf{y})$	–	Section 4.4.12
$-\sqrt{\alpha^2 - \|\mathbf{x}\|^2}$	$B[\mathbf{0}, \alpha]$	$\alpha\sqrt{\|\mathbf{y}\|_*^2 + 1}$	$\alpha > 0$	Section 4.4.13
$\sqrt{\alpha^2 + \|\mathbf{x}\|^2}$	$\mathbb{E}$	$-\alpha\sqrt{1 - \|\mathbf{y}\|_*^2}$ $(\mathrm{dom} f^* = B_{\|\cdot\|_*}[\mathbf{0}, 1])$	$\alpha > 0$	Section 4.4.14
$\frac{1}{2}\|\mathbf{x}\|^2$	$\mathbb{E}$	$\frac{1}{2}\|\mathbf{y}\|_*^2$	–	Section 4.4.15
$\frac{1}{2}\|\mathbf{x}\|^2 + \delta_C(\mathbf{x})$	C	$\frac{1}{2}\|\mathbf{y}\|^2 - \frac{1}{2}d_C^2(\mathbf{y})$	$\emptyset \neq C \subseteq \mathbb{E}, \mathbb{E}$ Euclidean	Example 4.4
$\frac{1}{2}\|\mathbf{x}\|^2 - \frac{1}{2}d_C^2(\mathbf{x})$	$\mathbb{E}$	$\frac{1}{2}\|\mathbf{y}\|^2 + \delta_C(\mathbf{y})$	$\emptyset \neq C \subseteq \mathbb{E}$ closed convex	Example 4.11

Conjugates of Symmetric Spectral Functions over $\mathbb{S}^n$ (From Example 7.16)

$g(\mathbf{X})$	$\mathrm{dom}(g)$	$g^*(\mathbf{Y})$	$\mathrm{dom}(g^*)$
$\lambda_{\max}(\mathbf{X})$	$\mathbb{S}^n$	$\delta_{\Upsilon_n}(\mathbf{Y})$	Υ_n
$\alpha\|\mathbf{X}\|_F \;(\alpha>0)$	$\mathbb{S}^n$	$\delta_{B_{\|\cdot\|_F}[\mathbf{0},\alpha]}(\mathbf{Y})$	$B_{\|\cdot\|_F}[\mathbf{0},\alpha]$
$\alpha\|\mathbf{X}\|_F^2 \;(\alpha>0)$	$\mathbb{S}^n$	$\frac{1}{4\alpha}\|\mathbf{Y}\|_F^2$	$\mathbb{S}^n$
$\alpha\|\mathbf{X}\|_{2,2} \;(\alpha>0)$	$\mathbb{S}^n$	$\delta_{B_{\|\cdot\|_{S_1}}[\mathbf{0},\alpha]}(\mathbf{Y})$	$B_{\|\cdot\|_{S_1}}[\mathbf{0},\alpha]$
$\alpha\|\mathbf{X}\|_{S_1} \;(\alpha>0)$	$\mathbb{S}^n$	$\delta_{B_{\|\cdot\|_{2,2}}[\mathbf{0},\alpha]}(\mathbf{Y})$	$B_{\|\cdot\|_{2,2}}[\mathbf{0},\alpha]$
$-\log\det(\mathbf{X})$	$\mathbb{S}^n_{++}$	$-n-\log\det(-\mathbf{Y})$	$\mathbb{S}^n_{--}$
$\displaystyle\sum_{i=1}^n \lambda_i(\mathbf{X})\log(\lambda_i(\mathbf{X}))$	$\mathbb{S}^n_+$	$\displaystyle\sum_{i=1}^n e^{\lambda_i(\mathbf{Y})-1}$	$\mathbb{S}^n$
$\displaystyle\sum_{i=1}^n \lambda_i(\mathbf{X})\log(\lambda_i(\mathbf{X}))$	Υ_n	$\log\left(\sum_{i=1}^n e^{\lambda_i(\mathbf{Y})}\right)$	$\mathbb{S}^n$

Conjugates of Symmetric Spectral Functions over $\mathbb{R}^{m\times n}$ (From Example 7.27)

$g(\mathbf{X})$	$\mathrm{dom}(g)$	$g^*(\mathbf{Y})$	$\mathrm{dom}(g^*)$
$\alpha\sigma_1(\mathbf{X}) \;(\alpha>0)$	$\mathbb{R}^{m\times n}$	$\delta_{B_{\|\cdot\|_{S_1}}[\mathbf{0},\alpha]}(\mathbf{Y})$	$B_{\|\cdot\|_{S_1}}[\mathbf{0},\alpha]$
$\alpha\|\mathbf{X}\|_F \;(\alpha>0)$	$\mathbb{R}^{m\times n}$	$\delta_{B_{\|\cdot\|_F}[\mathbf{0},\alpha]}(\mathbf{Y})$	$B_{\|\cdot\|_F}[\mathbf{0},\alpha]$
$\alpha\|\mathbf{X}\|_F^2 \;(\alpha>0)$	$\mathbb{R}^{m\times n}$	$\frac{1}{4\alpha}\|\mathbf{Y}\|_F^2$	$\mathbb{R}^{m\times n}$
$\alpha\|\mathbf{X}\|_{S_1} \;(\alpha>0)$	$\mathbb{R}^{m\times n}$	$\delta_{B_{\|\cdot\|_{S_\infty}}[\mathbf{0},\alpha]}(\mathbf{Y})$	$B_{\|\cdot\|_{S_\infty}}[\mathbf{0},\alpha]$

Smooth Functions

$f(\mathbf{x})$	$\mathrm{dom}(f)$	parameter	norm	reference
$\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}+\mathbf{b}^T\mathbf{x}+c$ $(\mathbf{A}\in\mathbb{S}^n, \mathbf{b}\in\mathbb{R}^n, c\in\mathbb{R})$	$\mathbb{R}^n$	$\|\mathbf{A}\|_{p,q}$	l_p	Example 5.2
$\langle\mathbf{b},\mathbf{x}\rangle+c$ $(\mathbf{b}\in\mathbb{E}^*, c\in\mathbb{R})$	$\mathbb{E}$	0	any norm	Example 5.3
$\frac{1}{2}\|\mathbf{x}\|_p^2, \;p\in[2,\infty)$	$\mathbb{R}^n$	$p-1$	l_p	Example 5.11
$\sqrt{1+\|\mathbf{x}\|_2^2}$	$\mathbb{R}^n$	1	l_2	Example 5.14
$\log(\sum_{i=1}^n e^{x_i})$	$\mathbb{R}^n$	1	l_2, l_∞	Example 5.15
$\frac{1}{2}d_C^2(\mathbf{x})$ $(C\subseteq\mathbb{E}\text{ closed convex})$	$\mathbb{E}$	1	Euclidean	Example 5.5
$\frac{1}{2}\|\mathbf{x}\|^2-\frac{1}{2}d_C^2(\mathbf{x})$ $(C\subseteq\mathbb{E}\text{ closed convex})$	$\mathbb{E}$	1	Euclidean	Example 5.6
$H_\mu(\mathbf{x}) \;(\mu>0)$	$\mathbb{E}$	$\frac{1}{\mu}$	Euclidean	Example 6.61

Strongly Convex Functions

$f(\mathbf{x})$	$\mathrm{dom}(f)$	s.c. parameter	norm	reference
$\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c$ $(\mathbf{A} \in \mathbb{S}^n_{++}, \mathbf{b} \in \mathbb{R}^n, c \in \mathbb{R})$	$\mathbb{R}^n$	$\lambda_{\min}(\mathbf{A})$	l_2	Example 5.19
$\frac{1}{2}\|\mathbf{x}\|^2 + \delta_C(\mathbf{x})$ $(\emptyset \neq C \subseteq \mathbb{E}\text{ convex})$	C	1	Euclidean	Example 5.21
$-\sqrt{1 - \|\mathbf{x}\|_2^2}$	$B_{\|\cdot\|_2}[\mathbf{0}, 1]$	1	l_2	Example 5.29
$\frac{1}{2}\|\mathbf{x}\|_p^2 \ \ (p \in (1, 2])$	$\mathbb{R}^n$	$p - 1$	l_p	Example 5.28
$\sum_{i=1}^n x_i \log x_i$	Δ_n	1	l_2 or l_1	Example 5.27

Orthogonal Projections

set (C)	$P_C(\mathbf{x})$	assumptions	reference
$\mathbb{R}^n_+$	$[\mathbf{x}]_+$	$-$	Lemma 6.26
$\mathrm{Box}[\boldsymbol{\ell}, \mathbf{u}]$	$P_C(\mathbf{x})_i = \min\{\max\{x_i, \ell_i\}, u_i\}$	$\ell_i \le u_i$	Lemma 6.26
$B_{\|\cdot\|_2}[\mathbf{c}, r]$	$\mathbf{c} + \frac{r}{\max\{\|\mathbf{x}-\mathbf{c}\|_2, r\}}(\mathbf{x} - \mathbf{c})$	$\mathbf{c} \in \mathbb{R}^n, r > 0$	Lemma 6.26
$\{\mathbf{x} : \mathbf{Ax} = \mathbf{b}\}$	$\mathbf{x} - \mathbf{A}^T(\mathbf{AA}^T)^{-1}(\mathbf{Ax} - \mathbf{b})$	$\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{A}$ full row rank	Lemma 6.26
$\{\mathbf{x} : \mathbf{a}^T\mathbf{x} \le b\}$	$\mathbf{x} - \frac{[\mathbf{a}^T\mathbf{x}-b]_+}{\|\mathbf{a}\|^2}\mathbf{a}$	$\mathbf{0} \ne \mathbf{a} \in \mathbb{R}^n, b \in \mathbb{R}$	Lemma 6.26
Δ_n	$[\mathbf{x} - \mu^*\mathbf{e}]_+$ where $\mu^* \in \mathbb{R}$ satisfies $\mathbf{e}^T[\mathbf{x} - \mu^*\mathbf{e}]_+ = 1$		Corollary 6.29
$H_{\mathbf{a},b} \cap \mathrm{Box}[\boldsymbol{\ell}, \mathbf{u}]$	$P_{\mathrm{Box}(\boldsymbol{\ell}, \mathbf{u})}(\mathbf{x} - \mu^*\mathbf{a})$ where $\mu^* \in \mathbb{R}$ satisfies $\mathbf{a}^T P_{\mathrm{Box}[\boldsymbol{\ell},\mathbf{u}]}(\mathbf{x} - \mu^*\mathbf{a}) = b$	$\mathbf{a} \in \mathbb{R}^n\backslash\{\mathbf{0}\}, b \in \mathbb{R}$	Theorem 6.27
$H^-_{\mathbf{a},b} \cap \mathrm{Box}[\boldsymbol{\ell}, \mathbf{u}]$	$\begin{cases} P_{\mathrm{Box}[\boldsymbol{\ell},\mathbf{u}]}(\mathbf{x}), & \mathbf{a}^T\mathbf{v_x} \le b, \\ P_{\mathrm{Box}[\boldsymbol{\ell},\mathbf{u}]}(\mathbf{x} - \lambda^*\mathbf{a}), & \mathbf{a}^T\mathbf{v_x} > b, \end{cases}$ $\mathbf{v_x} = P_{\mathrm{Box}[\boldsymbol{\ell},\mathbf{u}]}(\mathbf{x})$, $\mathbf{a}^T P_{\mathrm{Box}[\boldsymbol{\ell},\mathbf{u}]}(\mathbf{x} - \lambda^*\mathbf{a}) = b, \lambda^* > 0$	$\mathbf{a} \in \mathbb{R}^n\backslash\{\mathbf{0}\}, b \in \mathbb{R}$	Example 6.32
$B_{\|\cdot\|_1}[\mathbf{0}, \alpha]$	$\begin{cases} \mathbf{x}, & \|\mathbf{x}\|_1 \le \alpha, \\ \mathcal{T}_{\lambda^*}(\mathbf{x}), & \|\mathbf{x}\|_1 > \alpha, \end{cases}$ $\|\mathcal{T}_{\lambda^*}(\mathbf{x})\|_1 = \alpha, \lambda^* > 0$	$\alpha > 0$	Example 6.33
$\{\mathbf{x} : \boldsymbol{\omega}^T\|\mathbf{x}\| \le \beta,$ $-\boldsymbol{\alpha} \le \mathbf{x} \le \boldsymbol{\alpha}\}$	$\begin{cases} \mathbf{v_x}, & \boldsymbol{\omega}^T|\mathbf{v_x}| \le \beta, \\ \mathcal{S}_{\lambda^*\boldsymbol{\omega},\boldsymbol{\alpha}}(\mathbf{x}), & \boldsymbol{\omega}^T|\mathbf{v_x}| > \beta, \end{cases}$ $\mathbf{v_x} = P_{\mathrm{Box}[-\boldsymbol{\alpha},\boldsymbol{\alpha}]}(\mathbf{x})$, $\boldsymbol{\omega}^T|\mathcal{S}_{\lambda^*\boldsymbol{\omega},\boldsymbol{\alpha}}(\mathbf{x})| = \beta, \lambda^* > 0$	$\boldsymbol{\omega} \in \mathbb{R}^n_{++}$, $\boldsymbol{\alpha} \in [0, \infty]^n$, $\beta \in \mathbb{R}_{++}$	Example 6.34
$\{\mathbf{x} > \mathbf{0} : \Pi x_i \ge \alpha\}$	$\begin{cases} \mathbf{x}, & \mathbf{x} \in C, \\ \left(\frac{x_j + \sqrt{x_j^2 + 4\lambda^*}}{2}\right)^n_{j=1}, & \mathbf{x} \notin C, \end{cases}$ $\Pi_{j=1}^n \left((x_j + \sqrt{x_j^2 + 4\lambda^*})/2\right) = \alpha, \lambda^* > 0$	$\alpha > 0$	Example 6.35
$\{(\mathbf{x}, s) : \|\mathbf{x}\|_2 \le s\}$	$\left(\frac{\|\mathbf{x}\|_2 + s}{2\|\mathbf{x}\|_2}\mathbf{x}, \frac{\|\mathbf{x}\|_2 + s}{2}\right)$ if $\|\mathbf{x}\|_2 \ge |s|$ $(\mathbf{0}, 0)$ if $s < \|\mathbf{x}\|_2 < -s$, $(\mathbf{x}, s)$ if $\|\mathbf{x}\|_2 \le s$.		Example 6.37
$\{(\mathbf{x}, s) : \|\mathbf{x}\|_1 \le s\}$	$\begin{cases} (\mathbf{x}, s), & \|\mathbf{x}\|_1 \le s, \\ (\mathcal{T}_{\lambda^*}(\mathbf{x}), s + \lambda^*), & \|\mathbf{x}\|_1 > s, \end{cases}$ $\|\mathcal{T}_{\lambda^*}(\mathbf{x})\|_1 - \lambda^* - s = 0, \lambda^* > 0$		Example 6.38

Orthogonal Projections onto Symmetric Spectral Sets in $\mathbb{S}^n$

set (T)	$P_T(\mathbf{X})$	assumptions
$\mathbb{S}_+^n$	$\mathbf{U}\mathrm{diag}([\boldsymbol{\lambda}(\mathbf{X})]_+)\mathbf{U}^T$	–
$\{\mathbf{X} : \ell\mathbf{I} \preceq \mathbf{X} \preceq u\mathbf{I}\}$	$\mathbf{U}\mathrm{diag}(\mathbf{v})\mathbf{U}^T,$ $v_i = \min\{\max\{\lambda_i(\mathbf{X}),\ell\}, u\}$	$\ell \leq u$
$B_{\|\cdot\|_F}[\mathbf{0}, r]$	$\frac{r}{\max\{\|\mathbf{X}\|_F, r\}}\mathbf{X}$	$r > 0$
$\{\mathbf{X} : \mathrm{Tr}(\mathbf{X}) \leq b\}$	$\mathbf{U}\mathrm{diag}(\mathbf{v})\mathbf{U}^T,\ \mathbf{v} = \boldsymbol{\lambda}(\mathbf{X}) - \frac{[\mathbf{e}^T\boldsymbol{\lambda}(\mathbf{X})-b]_+}{n}\mathbf{e}$	$b \in \mathbb{R}$
Υ_n	$\mathbf{U}\mathrm{diag}(\mathbf{v})\mathbf{U}^T,\ \mathbf{v} = [\boldsymbol{\lambda}(\mathbf{X}) - \mu^*\mathbf{e}]_+$ where $\mu^* \in \mathbb{R}$ satisfies $\mathbf{e}^T[\boldsymbol{\lambda}(\mathbf{X}) - \mu^*\mathbf{e}]_+ = 1$	-
$B_{\|\cdot\|_{S_1}}[\mathbf{0}, \alpha]$	$\begin{cases} \mathbf{X}, & \|\mathbf{X}\|_{S_1} \leq \alpha, \\ \mathbf{U}\mathrm{diag}(\mathcal{T}_{\beta^*}(\boldsymbol{\lambda}(\mathbf{X})))\mathbf{U}^T, & \|\mathbf{X}\|_{S_1} > \alpha, \end{cases}$ $\|\mathcal{T}_{\beta^*}(\boldsymbol{\lambda}(\mathbf{X}))\|_1 = \alpha,\ \beta^* > 0$	$\alpha > 0$

Orthogonal Projections onto Symmetric Spectral Sets in $\mathbb{R}^{m \times n}$ (From Example 7.31)

set (T)	$P_T(\mathbf{X})$	assumptions
$B_{\|\cdot\|_{S_\infty}}[\mathbf{0}, \alpha]$	$\mathbf{U}\mathrm{diag}(\mathbf{v})\mathbf{U}^T,\ v_i = \min\{\sigma_i(\mathbf{X}), \alpha\}$	$\alpha > 0$
$B_{\|\cdot\|_F}[\mathbf{0}, r]$	$\frac{r}{\max\{\|\mathbf{X}\|_F, r\}}\mathbf{X}$	$r > 0$
$B_{\|\cdot\|_{S_1}}[\mathbf{0}, \alpha]$	$\begin{cases} \mathbf{X}, & \|\mathbf{X}\|_{S_1} \leq \alpha, \\ \mathbf{U}\mathrm{diag}(\mathcal{T}_{\beta^*}(\sigma(\mathbf{X})))\mathbf{V}^T, & \|\mathbf{X}\|_{S_1} > \alpha, \end{cases}$ $\|\mathcal{T}_{\beta^*}(\boldsymbol{\sigma}(\mathbf{X}))\|_1 = \alpha,\ \beta^* > 0$	$\alpha > 0$

Prox Calculus Rules

$f(\mathbf{x})$	$\mathrm{prox}_f(\mathbf{x})$	assumptions	reference
$\sum_{i=1}^m f_i(\mathbf{x}_i)$	$\mathrm{prox}_{f_1}(\mathbf{x}_1) \times \cdots \times \mathrm{prox}_{f_m}(\mathbf{x}_m)$		Theorem 6.6
$g(\lambda\mathbf{x} + \mathbf{a})$	$\frac{1}{\lambda}\left[\mathrm{prox}_{\lambda^2 g}(\mathbf{a} + \lambda\mathbf{x}) - \mathbf{a}\right]$	$\lambda \neq 0, \mathbf{a} \in \mathbb{E}, g$ proper	Theorem 6.11
$\lambda g(\mathbf{x}/\lambda)$	$\lambda\mathrm{prox}_{g/\lambda}(\mathbf{x}/\lambda)$	$\lambda \neq 0, g$ proper	Theorem 6.12
$g(\mathbf{x}) + \frac{c}{2}\|\mathbf{x}\|^2 + \langle\mathbf{a}, \mathbf{x}\rangle + \gamma$	$\mathrm{prox}_{\frac{1}{c+1}g}\left(\frac{\mathbf{x}-\mathbf{a}}{c+1}\right)$	$\mathbf{a} \in \mathbb{E}, c > 0, \gamma \in \mathbb{R},\ g$ proper	Theorem 6.13
$g(\mathcal{A}(\mathbf{x}) + \mathbf{b})$	$\mathbf{x} + \frac{1}{\alpha}\mathcal{A}^T(\mathrm{prox}_{\alpha g}(\mathcal{A}(\mathbf{x}) + \mathbf{b}) - \mathcal{A}(\mathbf{x}) - \mathbf{b})$	$\mathbf{b} \in \mathbb{R}^m,$ $\mathcal{A} : \mathbb{V} \to \mathbb{R}^m,$ g closed proper convex, $\mathcal{A} \circ \mathcal{A}^T = \alpha I,$ $\alpha > 0$	Theorem 6.15
$g(\|\mathbf{x}\|)$	$\mathrm{prox}_g(\|\mathbf{x}\|)\frac{\mathbf{x}}{\|\mathbf{x}\|},\quad \mathbf{x} \neq \mathbf{0}$ $\{\mathbf{u} : \|\mathbf{u}\| = \mathrm{prox}_g(0)\},\quad \mathbf{x} = \mathbf{0}$	g proper, $\mathrm{dom}(g) \subseteq [0, \infty)$	Theorem 6.18

Prox Computations

$f(\mathbf{x})$	$\mathrm{dom}(f)$	$\mathrm{prox}_f(\mathbf{x})$	assumptions	reference
$\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$	$\mathbb{R}^n$	$(\mathbf{A}+\mathbf{I})^{-1}(\mathbf{x}-\mathbf{b})$	$\mathbf{A} \in \mathbb{S}_+^n$, $\mathbf{b} \in \mathbb{R}^n$, $c \in \mathbb{R}$	Section 6.2.3
λx^3	$\mathbb{R}_+$	$\frac{-1+\sqrt{1+12\lambda[x]_+}}{6\lambda}$	$\lambda > 0$	Lemma 6.5
μx	$[0,\alpha]$	$\min\{\max\{x-\mu,0\},\alpha\}$	$\mu \in \mathbb{R}$, $\alpha \in \mathbb{R}_+$	Example 6.14
$\lambda\|\mathbf{x}\|$	$\mathbb{E}$	$\left(1 - \frac{\lambda}{\max\{\|\mathbf{x}\|,\lambda\}}\right)\mathbf{x}$	$\|\cdot\|$ - Euclidean norm, $\lambda > 0$	Example 6.19
$-\lambda\|\mathbf{x}\|$	$\mathbb{E}$	$\left(1 + \frac{\lambda}{\|\mathbf{x}\|}\right)\mathbf{x}, \qquad \mathbf{x} \neq \mathbf{0},$ $\{\mathbf{u} : \|\mathbf{u}\| = \lambda\}, \qquad \mathbf{x} = \mathbf{0}.$	$\|\cdot\|$ - Euclidean norm, $\lambda > 0$	Example 6.21
$\lambda\|\mathbf{x}\|_1$	$\mathbb{R}^n$	$\mathcal{T}_\lambda(\mathbf{x}) = [|\mathbf{x}| - \lambda\mathbf{e}]_+ \odot \mathrm{sgn}(\mathbf{x})$	$\lambda > 0$	Example 6.8
$\|\boldsymbol{\omega}\odot\mathbf{x}\|_1$	$\mathrm{Box}[-\boldsymbol{\alpha},\boldsymbol{\alpha}]$	$\mathcal{S}_{\boldsymbol{\omega},\boldsymbol{\alpha}}(\mathbf{x})$	$\boldsymbol{\alpha} \in [0,\infty]^n, \boldsymbol{\omega} \in \mathbb{R}_{++}^n$	Example 6.23
$\lambda\|\mathbf{x}\|_\infty$	$\mathbb{R}^n$	$\mathbf{x} - \lambda P_{B_{\|\cdot\|_1}[\mathbf{0},1]}(\mathbf{x}/\lambda)$	$\lambda > 0$	Example 6.48
$\lambda\|\mathbf{x}\|_a$	$\mathbb{E}$	$\mathbf{x} - \lambda P_{B_{\|\cdot\|_{a,*}}[\mathbf{0},1]}(\mathbf{x}/\lambda)$	$\|\mathbf{x}\|_a$ – norm, $\lambda > 0$	Example 6.47
$\lambda\|\mathbf{x}\|_0$	$\mathbb{R}^n$	$H_{\sqrt{2\lambda}}(x_1) \times \cdots \times H_{\sqrt{2\lambda}}(x_n)$	$\lambda > 0$	Example 6.10
$\lambda\|\mathbf{x}\|^3$	$\mathbb{E}$	$\frac{2}{1+\sqrt{1+12\lambda\|\mathbf{x}\|}}\mathbf{x}$	$\|\cdot\|$ - Euclidean norm, $\lambda > 0$,	Example 6.20
$-\lambda\sum_{j=1}^n \log x_j$	$\mathbb{R}_{++}^n$	$\left(\frac{x_j + \sqrt{x_j^2 + 4\lambda}}{2}\right)_{j=1}^n$	$\lambda > 0$	Example 6.9
$\delta_C(\mathbf{x})$	$\mathbb{E}$	$P_C(\mathbf{x})$	$\emptyset \neq C \subseteq \mathbb{E}$	Theorem 6.24
$\lambda\sigma_C(\mathbf{x})$	$\mathbb{E}$	$\mathbf{x} - \lambda P_C(\mathbf{x}/\lambda)$	$\lambda > 0$, $C \neq \emptyset$ closed convex	Theorem 6.46
$\lambda\max\{x_i\}$	$\mathbb{R}^n$	$\mathbf{x} - P_{\Delta_n}(\mathbf{x}/\lambda)$	$\lambda > 0$	Example 6.49
$\lambda\sum_{i=1}^k x_{[i]}$	$\mathbb{R}^n$	$\mathbf{x} - \lambda P_C(\mathbf{x}/\lambda),$ $C = H_{\mathbf{e},k} \cap \mathrm{Box}[\mathbf{0},\mathbf{e}]$	$\lambda > 0$	Example 6.50
$\lambda\sum_{i=1}^k |x_{\langle i\rangle}|$	$\mathbb{R}^n$	$\mathbf{x} - \lambda P_C(\mathbf{x}/\lambda),$ $C = B_{\|\cdot\|_1}[\mathbf{0},k] \cap \mathrm{Box}[-\mathbf{e},\mathbf{e}]$	$\lambda > 0$	Example 6.51
$\lambda M_f^\mu(\mathbf{x})$	$\mathbb{E}$	$\mathbf{x} + \frac{\lambda}{\mu+\lambda}\left(\mathrm{prox}_{(\mu+\lambda)f}(\mathbf{x}) - \mathbf{x}\right)$	$\lambda,\mu > 0$, f proper closed convex	Corollary 6.63
$\lambda d_C(\mathbf{x})$	$\mathbb{E}$	$\mathbf{x} + \min\left\{\frac{\lambda}{d_C(\mathbf{x})},1\right\}(P_C(\mathbf{x}) - \mathbf{x})$	C nonempty closed convex, $\lambda > 0$	Lemma 6.43
$\frac{\lambda}{2}d_C^2(\mathbf{x})$	$\mathbb{E}$	$\frac{\lambda}{\lambda+1}P_C(\mathbf{x}) + \frac{1}{\lambda+1}\mathbf{x}$	C nonempty closed convex, $\lambda > 0$	Example 6.64
$\lambda H_\mu(\mathbf{x})$	$\mathbb{E}$	$\left(1 - \frac{\lambda}{\max\{\|\mathbf{x}\|,\mu+\lambda\}}\right)$	$\lambda,\mu > 0$	Example 6.65
$\rho\|\mathbf{x}\|_1^2$	$\mathbb{R}^n$	$\left(\frac{v_i x_i}{v_i + 2\rho}\right)_{i=1}^n$, $\mathbf{v} = \left[\sqrt{\frac{\rho}{\mu}}|\mathbf{x}| - 2\rho\right]_+, \mathbf{e}^T\mathbf{v} = 1$ ($\mathbf{0}$ when $\mathbf{x} = \mathbf{0}$)	$\rho > 0$	Lemma 6.69
$\|\mathbf{A}\mathbf{x}\|_2$	$\mathbb{R}^n$	$\mathbf{x} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \alpha^*\mathbf{I})^{-1}\mathbf{A}\mathbf{x}$, $\alpha^* = 0$ if $\|\mathbf{v}_0\|_2 \leq \lambda$; otherwise, $\|\mathbf{v}_{\alpha^*}\|_2 = \lambda$; $\mathbf{v}_\alpha \equiv (\mathbf{A}\mathbf{A}^T + \alpha\mathbf{I})^{-1}\mathbf{A}\mathbf{x}$	$\mathbf{A} \in \mathbb{R}^{m\times n}$ with full row rank	Lemma 6.67

469

Prox of Symmetric Spectral Functions over $\mathbb{S}^n$ (From Example 7.19)

$F(\mathbf{X})$	$\mathrm{dom}(F)$	$\mathrm{prox}_F(\mathbf{X})$
$\alpha\|\mathbf{X}\|_F^2$	$\mathbb{S}^n$	$\frac{1}{1+2\alpha}\mathbf{X}$
$\alpha\|\mathbf{X}\|_F$	$\mathbb{S}^n$	$\left(1-\frac{\alpha}{\max\{\|\mathbf{X}\|_F,\alpha\}}\right)\mathbf{X}$
$\alpha\|\mathbf{X}\|_{S_1}$	$\mathbb{S}^n$	$\mathbf{U}\mathrm{diag}(\mathcal{T}_\alpha(\boldsymbol{\lambda}(\mathbf{X})))\mathbf{U}^T$
$\alpha\|\mathbf{X}\|_{2,2}$	$\mathbb{S}^n$	$\mathbf{U}\mathrm{diag}(\boldsymbol{\lambda}(\mathbf{X})-\alpha P_{B_{\|\cdot\|_1}[\mathbf{0},1]}(\boldsymbol{\lambda}(\mathbf{X})/\alpha))\mathbf{U}^T$
$-\alpha\log\det(\mathbf{X})$	$\mathbb{S}^n_{++}$	$\mathbf{U}\mathrm{diag}\left(\frac{\lambda_j(\mathbf{X})+\sqrt{\lambda_j(\mathbf{X})^2+4\alpha}}{2}\right)\mathbf{U}^T$
$\alpha\lambda_1(\mathbf{X})$	$\mathbb{S}^n$	$\mathbf{U}\mathrm{diag}(\boldsymbol{\lambda}(\mathbf{X})-\alpha P_{\Delta_n}(\boldsymbol{\lambda}(\mathbf{X})/\alpha))\mathbf{U}^T$
$\alpha\sum_{i=1}^k\lambda_i(\mathbf{X})$	$\mathbb{S}^n$	$\mathbf{X}-\alpha\mathbf{U}P_C(\boldsymbol{\lambda}(\mathbf{X})/\alpha)\mathbf{U}^T$, $C=H_{\mathbf{e},k}\cap\mathrm{Box}[\mathbf{0},\mathbf{e}]$

Prox of Symmetric Spectral Functions over $\mathbb{R}^{m\times n}$ (From Example 7.30)

$F(\mathbf{X})$	$\mathrm{prox}_F(\mathbf{X})$
$\alpha\|\mathbf{X}\|_F^2$	$\frac{1}{1+2\alpha}\mathbf{X}$
$\alpha\|\mathbf{X}\|_F$	$\left(1-\frac{\alpha}{\max\{\|\mathbf{X}\|_F,\alpha\}}\right)\mathbf{X}$
$\alpha\|\mathbf{X}\|_{S_1}$	$\mathbf{U}\mathrm{dg}(\mathcal{T}_\alpha(\boldsymbol{\sigma}(\mathbf{X})))\mathbf{V}^T$
$\alpha\|\mathbf{X}\|_{S_\infty}$	$\mathbf{X}-\alpha\mathbf{U}\mathrm{dg}(P_{B_{\|\cdot\|_1}[\mathbf{0},1]}(\boldsymbol{\sigma}(\mathbf{X})/\alpha))\mathbf{V}^T$
$\alpha\|\mathbf{X}\|_{\langle k\rangle}$	$\mathbf{X}-\alpha\mathbf{U}\mathrm{dg}(P_C(\boldsymbol{\sigma}(\mathbf{X})/\alpha))\mathbf{V}^T$,
	$C=B_{\|\cdot\|_1}[\mathbf{0},k]\cap B_{\|\cdot\|_\infty}[\mathbf{0},1]$