APG

DSA5103 Lecture 4

Yangjing Zhang

02-Feb-2023

NUS

Today's content

1. Basic convex analysis
2. Proximal operator
3. (Accelerated) proximal gradient method

Basic convex analysis

Norms

A vector norm on $\mathbb{R}^n$ is a function $\|\cdot\| : \mathbb{R} \to \mathbb{R}$ satisfying the following properties:

(1) $\|x\| \geq 0 \quad \forall\, x \in \mathbb{R}^n$, and $\|x\| = 0 \iff x = 0$

(2) $\|\alpha x\| = |\alpha| \|x\| \quad \forall\, \alpha \in \mathbb{R},\, x \in \mathbb{R}^n$

(3) $\|x + y\| \leq \|x\| + \|y\| \quad \forall\, x, y \in \mathbb{R}^n$

Example.

1. $\|x\|_1 = \sum_{i=1}^n |x_i|$ (ℓ_1 norm)
2. $\|x\|_2 = (\sum_{i=1}^n |x_i|^2)^{1/2}$ (ℓ_2 norm)
3. $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$, where $1 \leq p < \infty$ (ℓ_p norm)
4. $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$ (ℓ_∞ norm)
5. $\|x\|_{W,p} = \|Wx\|_p$, where W is a fixed nonsingular matrix, $1 \leq p \leq \infty$

Inner product

For the space of $m \times n$ matrices, $\mathbb{R}^{m \times n}$, we define the standard inner product, for any $A, B \in \mathbb{R}^{m \times n}$,

$$\langle A, B \rangle = \mathrm{Tr}(A^T B) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij}$$

Recap: the trace of a square matrix $C \in \mathbb{R}^{n \times n}$ is $\mathrm{Tr}(C) = \sum_{i=1}^{n} C_{ii}$.

For the space of n-vectors, $\mathbb{R}^n$, we define the standard inner product, for any $x, y \in \mathbb{R}^n$,

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

Projection onto a closed convex set

Theorem (Projection theorem). Let C be a closed convex set.

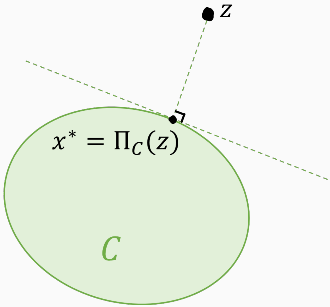
(1) For every z, there exists a unique minimizer of

$$\min_{x \in C} \quad \frac{1}{2}\|x - z\|^2,$$

denoted as $\Pi_C(z)$ and called as the projection of z onto C.

(2) $x^* := \Pi_C(z)$ is the projection of z onto C if and only if

$$\langle z - x^*, x - x^* \rangle \leq 0 \quad \forall\, x \in C.$$



Projection onto a closed convex set

Theorem (Projection theorem). Let C be a closed convex set.

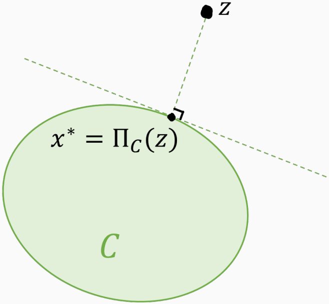
(1) For every z, there exists a unique minimizer of

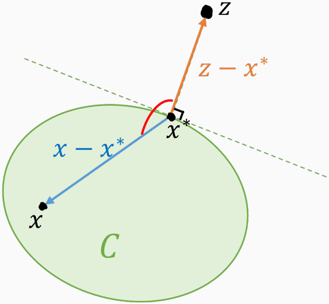
$$\min_{x \in C} \quad \frac{1}{2}\|x - z\|^2,$$

denoted as $\Pi_C(z)$ and called as the projection of z onto C.

(2) $x^* := \Pi_C(z)$ is the projection of z onto C if and only if

$$\langle z - x^*, x - x^* \rangle \leq 0 \quad \forall\, x \in C.$$





Projection onto a closed convex set

Example.

1. $C = \mathbb{R}^n_+ = \{x \in \mathbb{R}^n \mid x_i \geq 0, \forall\, i = 1, 2, \ldots, n\}$ positive orthant

$$\Pi_C(z) = \Pi_{\mathbb{R}^n_+}(z) = \max\{z, 0\}$$

2. $C = \{x \in \mathbb{R}^n \mid \|x\|_p \leq 1\}$ ℓ_p ball

$$\Pi_C(z) = \begin{cases} z, & \text{if } \|z\|_p \leq 1 \\ \frac{z}{\|z\|_p}, & \text{if } \|z\|_p > 1 \end{cases} = \frac{z}{\max\{\|z\|_p, 1\}}$$

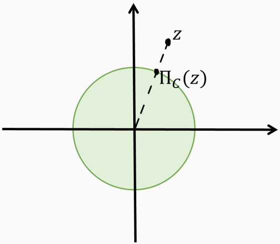


Figure 1: $C = \{x \in \mathbb{R}^2 \mid \|x\|_2 \leq 1\}$ ℓ_2 ball

Projection onto a closed convex set

Example.

3. $C = \mathbb{S}^n_+$ the space of $n \times n$ symmetric and positive semidefinite matrices

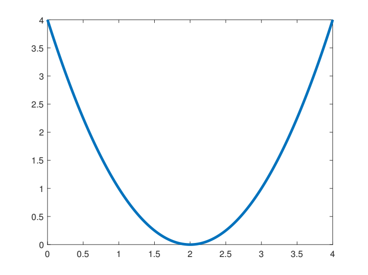
$$\Pi_C(A) = \Pi_{\mathbb{S}^n_+}(A) = Q \begin{bmatrix} \max\{\lambda_1, 0\} & & \\ & \ddots & \\ & & \max\{\lambda_n, 0\} \end{bmatrix} Q^T$$

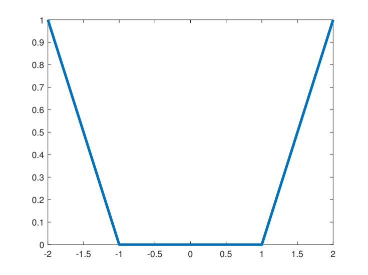
for a given $A \in \mathbb{S}^n$ with eigenvalue decomposition

$$A = Q \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} Q^T$$

arg min

$\arg\min_x f(x)$ denotes the solution set of x for which $f(x)$ attains its minimum (argument of the minimum)





Left: $\min_x f(x) = 0$, $\arg\min_x f(x) = \{2\}$

Right: $\min_x f(x) = 0$, $\arg\min_x f(x) = [-1, 1]$

$\Pi_C(z) = \arg\min_{x\in C} \quad \frac{1}{2}\|x - z\|^2$

Extended real-valued function

Definition.

Let $\mathcal{X}$ be a Euclidean space (e.g., $\mathcal{X} = \mathbb{R}^n$ or $\mathbb{R}^{m\times n}$). Let $f : \mathcal{X} \to (-\infty, +\infty]$ be an extended real-valued function[1].

(1) The (effective) domain of f is defined to be the set

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

(2) f is said to be proper if $\mathrm{dom}(f) \neq \emptyset$.

(3) f is said to be closed if its epi-graph

$$\mathrm{epi}(f) := \{(x, \alpha) \in \mathcal{X} \times \mathbb{R} \mid f(x) \leq \alpha\}$$

is closed.

(4) f is said to be convex if its epi-graph is convex.

[1] Here f is allowed to take the value of $+\infty$, but not allowed to take the value of $-\infty$

Extended real-valued function

- For a real-valued function $f : \mathcal{X} \to \mathbb{R}$, this definition of convexity

$$\mathrm{epi}(f) \text{ is convex} \tag{1}$$

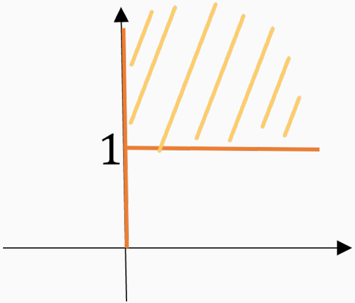
  coincides with the one we have used

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y),\ \forall\, x, y \in \mathrm{dom}(f),\ \lambda \in [0,1] \tag{2}$$

  [Exercise] (1) $\iff$ (2)
- A convex function $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ can be extended to a convex function on all of $\mathbb{R}^n$ by setting $f(x) = +\infty$ for $x \notin D$.

Example: extended real-valued function

$$f(x) = \begin{cases} 0, & \text{if } x = 0 \\ 1, & \text{if } x > 0 \\ +\infty, & \text{if } x < 0 \end{cases}$$

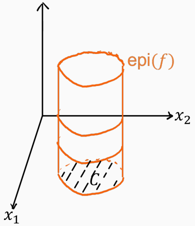


- $\mathrm{dom}(f) = [0, +\infty)$, f is proper
- $\mathrm{epi}(f) = \{0\} \times [0, +\infty) \cup (0, +\infty) \times [1, +\infty)$ is closed, i.e., f is closed
- $\mathrm{epi}(f)$ is not convex, i.e., f is not convex

Indicator/support function

Let C be a nonempty set in $\mathcal{X}$.

$f : \mathbb{R}^2 \to (-\infty, +\infty]$



The indicator function of C is

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

- $\mathrm{dom}(f) = C$, f is proper
- $\mathrm{epi}(f) = C \times [0, +\infty)$ is closed if C is closed, i.e., $\delta_C(\cdot)$ is closed if C is closed
- $\mathrm{epi}(f)$ is convex if C is convex, i.e., $\delta_C(\cdot)$ is convex if C is convex

The support function of C is

$$\delta_C^*(x) = \max_{y \in C} \quad \langle x, y \rangle.$$

Indicator and support functions give correspondences between convex sets and convex functions.

Dual/polar cone

Definition (Cone)

A set $C \subseteq \mathcal{X}$ is called a cone if $\lambda x \in C$ when $x \in C$ and $\lambda \geq 0$.

Definition (Dual and polar cone)

The dual cone of a set $C \subseteq \mathcal{X}$ (not necessarily convex) is defined by

$$C^* = \{y \in \mathcal{X} \mid \langle x, y \rangle \geq 0 \quad \forall x \in C\}$$

The polar cone of C is $C^o = -C^*$.

If $C^* = C$, then C is said to be self-dual.

- C^* is always a convex cone, even if C is neither convex nor a cone.

Dual/polar cone

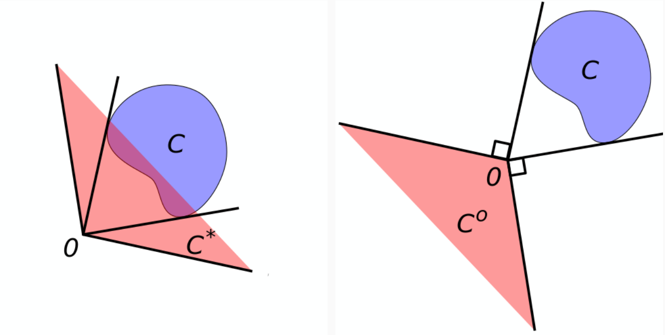


Figure 2: Left: A set C and its dual cone C^*. Right: A set C and its polar cone C^o. The dual cone and the polar cone are symmetric to each other with respect to the origin. Image from internet

Example: self-dual cones

1. $\mathcal{X} = \mathbb{R}^n$. $C = \mathbb{R}^n_+$ is a self-dual closed convex cone
   - ▷ Proof: $C^* = \{y \in \mathbb{R}^n \mid \langle x, y\rangle \geq 0 \quad \forall x \in \mathbb{R}^n_+\} = \mathbb{R}^n_+$

2. $\mathcal{X} = \mathbb{S}^n$. $C = \mathbb{S}^n_+$ is a self-dual closed convex cone (psd cone)
   - ▷ Proof: want to show that $\{B \in \mathbb{S}^n \mid \langle A, B\rangle \geq 0 \quad \forall A \in \mathbb{S}^n_+\} = \mathbb{S}^n_+$
     [LHS $\subseteq$ RHS] Take $B \in C^*$. For any $x \in \mathbb{R}^n$, we have $xx^T \in \mathbb{S}^n_+$ and thus $\langle xx^T, B\rangle = x^T B x \geq 0$, which implies $B \in \mathbb{S}^n_+$.
     [RHS $\subseteq$ LHS] Take $B \in \mathbb{S}^n_+$. Compute its eigenvalue decomposition $B = \sum_{i=1}^n \lambda_i v_i v_i^T$, $\lambda_i \geq 0$. For any $A \in \mathbb{S}^n_+$, we have $\langle A, B\rangle = \langle A, \sum_{i=1}^n \lambda_i v_i v_i^T\rangle = \sum_{i=1}^n \lambda_i v_i^T A v_i \geq 0$.

Example: self-dual cones

3. $\mathcal{X} = \mathbb{R}^n$. $C := \{x \in \mathbb{R}^n \mid \sqrt{x_2^2 + \cdots + x_n^2} \leq x_1, x_1 \geq 0\}$ is a self-dual closed convex cone (second-order cone)
   - ▷ Proof: want to show that $\{y \in \mathbb{R}^n \mid \langle x, y \rangle \geq 0 \quad \forall x \in C\} = C$
     [RHS $\subseteq$ LHS] Take $y \in C$. For any $x \in C$

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n \geq x_1 y_1 - \sqrt{\sum_{i=2}^n x_i^2} \sqrt{\sum_{i=2}^n y_i^2} \geq 0$$

The first inequality follows from the Cauchy-Schwartz inequality, and the last inequality follows from the fact that $x, y \in C$.
[LHS $\subseteq$ RHS] Take $y \in C^*$. If $[y_2; \ldots; y_n] = 0$, we take $x = [1; 0; \ldots; 0] \in C$ then $\langle x, y \rangle = y_1 \geq 0$. Obviously, $y \in C$. Else, we take

$$x = \left[\sqrt{\sum_{i=2}^n y_i^2}; -y_2^2; \ldots; -y_n^2 \right] \in C$$

then

$$\langle x, y \rangle = y_1 \sqrt{\sum_{i=2}^n y_i^2} - y_2^2 - y_n^2 \geq 0 \Rightarrow y_1 \geq \sqrt{\sum_{i=2}^n y_i^2} \Rightarrow y \in C$$

Example: self-dual cones

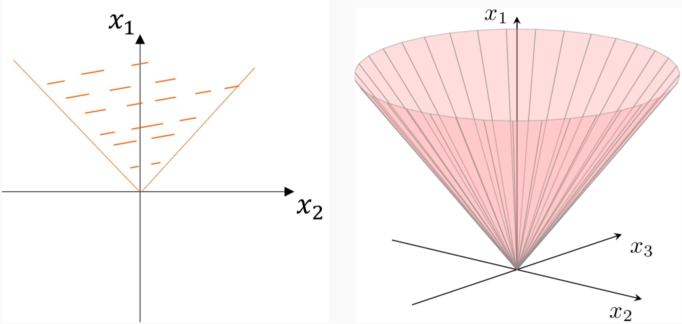


Figure 3: Left: second-order cone $\{x \in \mathbb{R}^2 \mid x_1 \geq |x_2|\}$. Right: second-order cone $\{x \in \mathbb{R}^3 \mid x_1 \geq \sqrt{x_2^2 + x_3^2}\}$. Image from internet

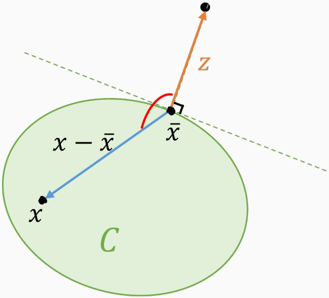
Normal cone

Definition (Normal cone)

Let C be a convex set in $\mathcal{X}$ and $\bar{x} \in C$. The normal cone of C at $\bar{x} \in C$ is defined by

$$N_C(\bar{x}) := \{z \in \mathcal{X} \mid \langle z, x - \bar{x} \rangle \geq 0 \quad \forall\, x \in C\}$$

By convention, we let $N_C(\bar{x}) = \emptyset$ if $\bar{x} \notin C$.



Example: normal cones

$$N_C(\bar{x}) := \{z \in \mathcal{X} \mid \langle z, x - \bar{x}\rangle \geq 0 \quad \forall\, x \in C\}$$

Example 1. $C = [0,1] \subseteq \mathbb{R}$

$$N_C(\bar{x}) = \begin{cases} (-\infty, 0], & \text{if } \bar{x} = 0 \\ [0, +\infty), & \text{if } \bar{x} = 1 \\ \{0\}, & \text{if } \bar{x} \in (0,1) \\ \emptyset, & \text{if } \bar{x} \notin C \end{cases}$$

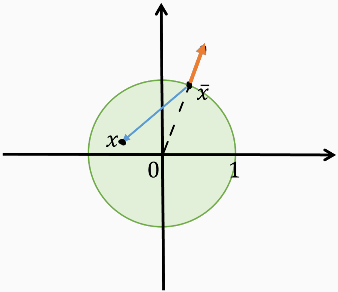
Example: normal cones

$$N_C(\bar{x}) := \{z \in \mathcal{X} \mid \langle z, x - \bar{x}\rangle \geq 0 \quad \forall\, x \in C\}$$

Example 2.

$C = \{x \in \mathbb{R}^2 \mid \|x\| \leq 1\} \subseteq \mathbb{R}^2$

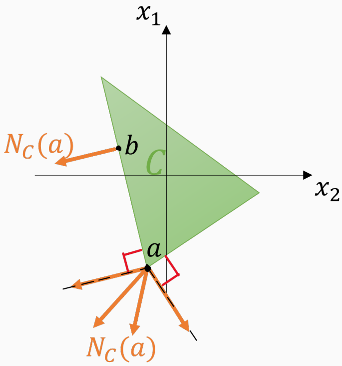
$$N_C(\bar{x}) = \begin{cases} \{\lambda \bar{x} \mid \lambda \geq 0\}, & \text{if } \|\bar{x}\| = 1 \\ \{0\}, & \text{if } \|\bar{x}\| < 1 \\ \emptyset, & \text{if } \bar{x} \notin C \end{cases}$$



Example: normal cones

$$N_C(\bar{x}) := \{z \in \mathcal{X} \mid \langle z, x - \bar{x} \rangle \geq 0 \quad \forall\, x \in C\}$$

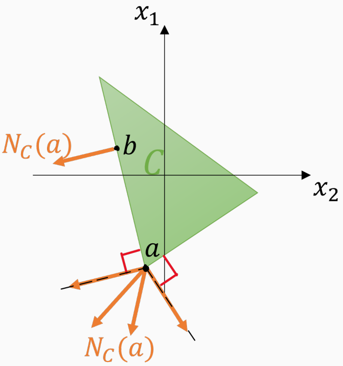
Example 3. $C =$ a triangle $\subseteq \mathbb{R}^2$



Example: normal cones

$$N_C(\bar{x}) := \{z \in \mathcal{X} \mid \langle z, x - \bar{x} \rangle \geq 0 \quad \forall\, x \in C\}$$

Example 3. $C =$ a triangle $\subseteq \mathbb{R}^2$



[Exercise] (1) $C = \{x \in \mathbb{R}^2 \mid x_1 + x_2 \leq 1, x_1 \geq 0, x_2 \geq 0\}$. Find $N_C(\bar{x})$ for $\bar{x} = [0; 1]$, $\bar{x} = [0.5; 0.5]$, $\bar{x} = [0.1; 0.2]$.

(2) $C = \mathbb{R}^n_+$. Find $N_C(\bar{x})$, $\bar{x} = [1; 1; 0; 0; \ldots; 0]$.

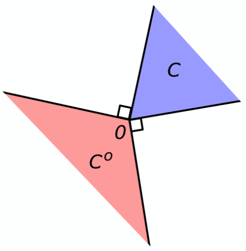
Normal cone

Proposition Let $C \subseteq \mathcal{X}$ be a nonempty convex set and $\bar{x} \in C$. Then

(1) $N_C(\bar{x})$ is a closed convex cone.

(2) If $\bar{x} \in \mathrm{int}(C)$ ($\bar{x}$ is an interior point of C), then $N_C(\bar{x}) = \{0\}$.

(3) If C is a cone, then $N_C(\bar{x}) \subseteq C^o$.



Normal cone

Proposition Let $C \subseteq \mathcal{X}$ be a nonempty closed convex set. Then for any $u, y \in C$,

$$u \in N_C(y) \iff y = \Pi_C(y+u)$$

Proof. "$\Rightarrow$" Suppose $u \in N_C(y)$. Then $\langle u, x-y \rangle \leq 0$ for all $x \in C$. Thus

$$\langle (y+u) - y, x-y \rangle \leq 0 \text{ for all } x \in C$$

which implies that $y = \Pi_C(y+u)$.

"$\Leftarrow$" Suppose $y = \Pi_C(y+u)$. Then we know that

$$\langle u, x-y \rangle = \langle (y+u) - y, x-y \rangle \leq 0 \text{ for all } x \in C$$

which implies that $u \in N_C(y)$.

Subdifferantial

Definition Let $f : \mathcal{X} \to (-\infty, +\infty]$ be a convex function.

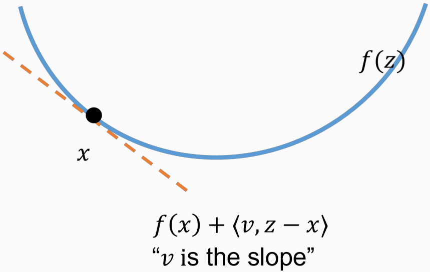
We call v a subgradient of f at $x \in \mathrm{dom}(f)$ if

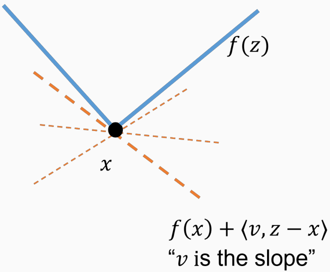
$$f(z) \geq f(x) + \langle v, z - x \rangle \quad \forall\, z \in \mathcal{X}.$$

The set of all subgradients at x is called the subdifferential of f at x, denoted as

$$\partial f(x) = \{v \mid f(z) \geq f(x) + \langle v, z - x \rangle \quad \forall\, z \in \mathcal{X}\}.$$

By convention, $\partial f(x) = \emptyset$ for any $x \notin \mathrm{dom}(f)$.





Subgradient is an extension of gradient

- If f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$.
  - ▷ Proof: If $v \in \partial f(x)$, then $f(x+h) \geq f(x) + \langle v, h \rangle \, \forall \, h$. By taking $h = t(v - \nabla f(x)), t > 0$, and use first-order Taylor series expansion to get $t\|v - \nabla f(x)\| \leq o(t) \, \forall \, t > 0$, which implies $v = \nabla f(x)$.

Theorem Let $f : \mathcal{X} \to (-\infty, +\infty]$ be a proper convex function. Then $\bar{x} \in \mathcal{X}$ is a global minimizer of $\min_{x \in \mathcal{X}} f(x)$ if and only if $0 \in \partial f(\bar{x})$.

Proof. By the subgradient inequality $f(z) \geq f(\bar{x}) + \langle v, z - \bar{x} \rangle \quad \forall \, z \in \mathcal{X}$.

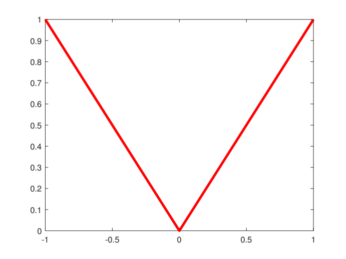
Take $0 = v \in \partial f(\bar{x})$

Example: subdifferantial

Example 1.

$f(x) = |x|$, $x \in \mathbb{R}$.

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

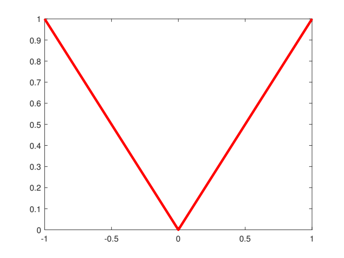


Example: subdifferantial

Example 1.

$f(x) = |x|$, $x \in \mathbb{R}$.

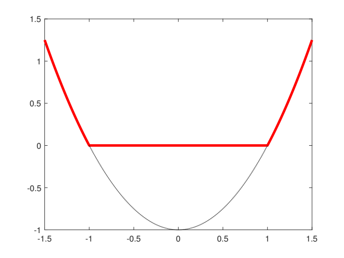
$$\partial f(x) = \begin{cases} \{-1\}, & \text{if } x < 0 \\ [-1, 1], & \text{if } x = 0 \\ \{1\}, & \text{if } x > 0 \end{cases}$$



Example 2.

$f(x) = \max\{x^2 - 1, 0\}$, $x \in \mathbb{R}$.

$$\partial f(x) = \begin{cases} \{2x\}, & \text{if } x < -1, x > 1 \\ \{0\}, & \text{if } -1 < x < 1 \\ [-2, 0], & \text{if } x = -1 \\ [0, 2], & \text{if } x = 1 \end{cases}$$



Example: subdifferantial

Example 3. Let C be a convex set.

$$\partial\delta_C(x) = \begin{cases} N_C(x), & \text{if } x \in C \\ \emptyset, & \text{if } x \notin C. \end{cases}$$

Proof: Take $x \in C$.

$$\begin{aligned} & v \in \partial\delta_C(x) \\ \iff & \delta_C(z) \geq \delta_C(x) + \langle v, z - x\rangle \quad \forall\, z \in C \\ \iff & 0 \geq \langle v, z - x\rangle \quad \forall\, z \in C \\ \iff & v \in N_C(x) \end{aligned}$$

Example: subdifferantial

Example 3. Let C be a convex set.

$$\partial\delta_C(x) = \begin{cases} N_C(x), & \text{if } x \in C \\ \emptyset, & \text{if } x \notin C. \end{cases}$$

Proof: Take $x \in C$.

$$\begin{aligned} & v \in \partial\delta_C(x) \\ \iff & \delta_C(z) \geq \delta_C(x) + \langle v, z - x\rangle \quad \forall\, z \in C \\ \iff & 0 \geq \langle v, z - x\rangle \quad \forall\, z \in C \\ \iff & v \in N_C(x) \end{aligned}$$

Example 4. $f(x) = \|x\|_1$, $x \in \mathbb{R}^n$. [Exercise] Show that

$$\partial f(0) = \{y \in \mathbb{R}^n \mid \|y\|_\infty \leq 1\}.$$

Lipschitz continuous

Definition (Lipschitz continuous)

A function $F : \mathbb{R}^n \to \mathbb{R}^m$ is said to be locally Lipschitz continuous if for any open set $\mathcal{O} \subseteq \mathbb{R}^n$, there exists a constant L (depending on $\mathcal{O}$) such that

$$\|F(x) - F(y)\| \leq L\|x - y\| \quad \forall\, x, y \in \mathcal{O}.$$

If $\mathcal{O} = \mathbb{R}^n$, then F is said to be globally Lipschitz continuous.

Example

(1) $f(x) = |x|, x \in \mathbb{R}$ is globally Lipschitz continuous with Lipschitz constant $L = 1$

(2) $f(x) = x^2, x \in \mathbb{R}$ is locally Lipschitz continuous but not globally Lipschitz continuous

Fenchel conjugate

Definition Let $f : \mathcal{X} \to [-\infty, +\infty]$. The (Fenchel) conjugate of f is defined by

$$f^*(y) = \sup\{\langle y, x\rangle - f(x) \mid x \in \mathcal{X}\},\ y \in \mathcal{X}$$

Remark

- f^* is always closed and convex, even if f is neither convex nor closed.
- If $f : \mathcal{X} \to (-\infty, +\infty]$ a closed proper convex function, then $(f^*)^* = f$.

Fenchel conjugate

Proposition Let $f : \mathcal{X} \to (-\infty, +\infty]$ be a closed proper convex function. The following is equivalent:

(1) $f(x) + f^*(y) = \langle x, y \rangle$

(2) $y \in \partial f(x)$

(3) $x \in \partial f^*(y)$

- "$y \in \partial f(x) \iff x \in \partial f^*(y)$" means that ∂f^* is the inverse of ∂f in the sense of multi-valued mappings.

Example: conjugate function

Example 1. Let $C \subseteq \mathcal{X}$ be a nonempty convex set. Compute the conjugate of the indicator function of C.

Solution. Recall that the indicator function of C is

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

Its conjugate

$$\delta_C^*(y) = \sup\{\langle y, x\rangle - \delta_C(x) \mid x \in \mathcal{X}\} = \sup\{\langle y, x\rangle \mid x \in C\}$$

is the support function. This explains why we use the notation δ_C^* for the support function of C.

[Exercise] Let $C \subseteq \mathcal{X}$ be a nonempty convex cone. Show that

$$\delta_C^*(y) = \delta_{C^\circ}(y) \quad \forall\, y$$

Example: conjugate function

Example 2. Let $f(x) = \|x\|_1$, $x \in \mathbb{R}^n$. Compute f^*.

Solution. $f^*(y) = \sup_x \{\langle y, x\rangle - \|x\|_1\}$

Case 1: if $\|y\|_\infty \leq 1$, we have $\langle y, x\rangle - \|x\|_1 \leq \|x\|_1 \|y\|_\infty - \|x\|_1 \leq 0$. Therefore, $f^*(y) = 0$.

Case 2: if $\|y\|_\infty > 1$, there exists $|y_k| > 1$. For a positive integer m, we construct

$$\bar{x} = [0; \ldots; 0; m\,\mathrm{sign}(y_k); 0; \ldots; 0]$$

and therefore $f^*(y) \geq \langle y, \bar{x}\rangle - \|\bar{x}\|_1 = m(|y_k| - 1) \to +\infty$, as $m \to +\infty$. Therefore, $f^*(y) = +\infty$.

In conclusion, $f^*(y) = \delta_C(y)$, $C = \{y \in \mathbb{R}^n \mid \|y\|_\infty \leq 1\}$.

Example: conjugate function

Example 2. Let $f(x) = \|x\|_1$, $x \in \mathbb{R}^n$. Compute f^*.

Solution. $f^*(y) = \sup_x\{\langle y, x\rangle - \|x\|_1\}$

Case 1: if $\|y\|_\infty \leq 1$, we have $\langle y, x\rangle - \|x\|_1 \leq \|x\|_1\|y\|_\infty - \|x\|_1 \leq 0$. Therefore, $f^*(y) = 0$.

Case 2: if $\|y\|_\infty > 1$, there exists $|y_k| > 1$. For a positive integer m, we construct

$$\bar{x} = [0; \ldots ; 0; m\,\mathrm{sign}(y_k); 0; \ldots ; 0]$$

and therefore $f^*(y) \geq \langle y, \bar{x}\rangle - \|\bar{x}\|_1 = m(|y_k| - 1) \to +\infty$, as $m \to +\infty$. Therefore, $f^*(y) = +\infty$.

In conclusion, $f^*(y) = \delta_C(y)$, $C = \{y \in \mathbb{R}^n \mid \|y\|_\infty \leq 1\}$.

[Exercise] Let $f(x) = \lambda\|x\|_p$, $x \in \mathbb{R}^n$, $1 < p < \infty$, $\lambda > 0$. Show that $f^*(y) = \delta_C(y)$, $C = \{y \in \mathbb{R}^n \mid \|y\|_q \leq \lambda\}$, $\frac{1}{p} + \frac{1}{q} = 1$.

Proximal operator

Moreau envelope and proximal mapping

Let $f : \mathcal{X} \to (-\infty, \infty]$ be a closed proper convex function. We define

- Moreau envelope (Moreau-Yosida regularization) of f at x

$$M_f(x) = \min_y \left\{ f(y) + \frac{1}{2}\|y - x\|^2 \right\}$$

- Proximal mapping of f at x

$$P_f(x) = \arg\min_y \left\{ f(y) + \frac{1}{2}\|y - x\|^2 \right\}$$

- $M_f(x)$ is differentiable, its gradient is $\nabla M_f(x) = x - P_f(x)$ (Moreau envelope is a way to smooth a possibly non-differentiable convex function)
- $P_f(x)$ exists and is unique
- $M_f(x) \leq f(x)$
- $\arg\min\ f(x) = \arg\min\ M_f(x)$

Example

Let $C \subseteq \mathcal{X}$ be a nonempty closed convex set and $f(x) = \delta_C(x)$ be the indicator function of C. Its proximal mapping is

$$P_f(x) = \arg\min_{y \in \mathcal{X}} \left\{ \delta_C(y) + \frac{1}{2}\|y - x\|^2 \right\} = \arg\min_{y \in C} \frac{1}{2}\|y - x\|^2 = \Pi_C(x)$$

Its Moreau envelope is

$$M_f(x) = \frac{1}{2}\|x - \Pi_C(x)\|^2$$

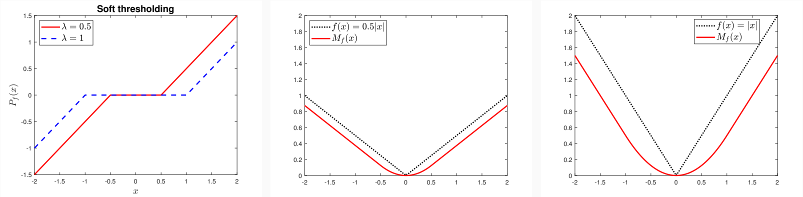
Example

Let $f(x) = \lambda|x|$, $x \in \mathbb{R}$. Its Moreau envelope (known as Huber function)

$$M_f(x) = \begin{cases} \frac{1}{2}x^2, & |x| \le \lambda \\ \lambda|x| - \frac{\lambda^2}{2}, & |x| > \lambda \end{cases}$$

Its proximal mapping is (known as soft thresholding)

$$P_f(x) = \mathrm{sign}(x)\max\{|x| - \lambda, 0\}$$



We can see that M_f is a smoothing of f, $M_f \le f$,
$\arg\min\ f(x) = \arg\min\ M_f(x)$.

Soft thresholding

The soft thresholding operator $S : \mathbb{R}^n \to \mathbb{R}^n$ is defined as

$$S_\lambda(x) = \begin{bmatrix} \mathrm{sign}(x_1) \max\{|x_1| - \lambda, 0\} \\ \mathrm{sign}(x_2) \max\{|x_2| - \lambda, 0\} \\ \vdots \\ \mathrm{sign}(x_n) \max\{|x_n| - \lambda, 0\} \end{bmatrix}$$

for any $x = [x_1; \ldots; x_n] \in \mathbb{R}^n$ and $\lambda > 0$.

Example. Given

$$x = \begin{bmatrix} 1.5 \\ -0.4 \\ 3 \\ -2 \\ 0.8 \end{bmatrix}, \quad S_{0.5}(x) = \begin{bmatrix} 1 \\ 0 \\ 2.5 \\ -1.5 \\ 0.3 \end{bmatrix}, \quad S_2(x) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Moreau envelope and proximal mapping

Theorem (Moreau decomposition)

Let $f : \mathcal{X} \to (-\infty, \infty]$ be a closed proper convex function and f^* be its conjugate. For any $x \in \mathcal{X}$, it holds that

$$\begin{aligned} x =& P_f(x) + P_{f^*}(x) \\ \frac{1}{2}\|x\|^2 =& M_f(x) + M_{f^*}(x) \end{aligned}$$

Example. Let $C \subseteq \mathcal{X}$ be a nonempty closed convex cone. $f(x) = \delta_C(x)$ and $f^*(x) = \delta_C^*(x) = \delta_{C^\circ}(x)$. Therefore

$$x = \Pi_C(x) + \Pi_{C^\circ}(x).$$

Moreau envelope and proximal mapping

Theorem (Moreau decomposition)

Let $f : \mathcal{X} \to (-\infty, \infty]$ be a closed proper convex function and f^* be its conjugate. For any $x \in \mathcal{X}$, it holds that

$$\begin{aligned} x =& P_f(x) + P_{f^*}(x) \\ \frac{1}{2}\|x\|^2 =& M_f(x) + M_{f^*}(x) \end{aligned}$$

Example. Let $C \subseteq \mathcal{X}$ be a nonempty closed convex cone. $f(x) = \delta_C(x)$ and $f^*(x) = \delta_C^*(x) = \delta_{C^o}(x)$. Therefore

$$x = \Pi_C(x) + \Pi_{C^o}(x).$$

- $M_f(\cdot)$ is always differentiable even though f is non-differentiable
- $P_f(\cdot)$ is important in many optimization algorithms (e.g., accelerated proximal gradient methods introduced later)
- For many widely used regularizaters, $P_f(\cdot)$ and $M_f(\cdot)$ have explicit expression

$$u \in N_C(y) \Updownarrow y = \Pi_C(y+u)$$

$$\begin{tikzcd}
\Pi_C \arrow[rr, leftrightarrow, "P_{\delta_C} = \Pi_C"] \arrow[d, leftrightarrow] & & \delta_C \arrow[d, leftrightarrow, "\text{conjugate}"] \arrow[dll, leftrightarrow, "\partial \delta_C = N_C"] \\
N_C & & \delta_C^*
\end{tikzcd}$$

(Accelerated) proximal gradient method

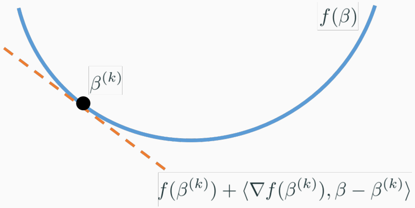
A proximal point view of gradient methods

To minimize a differentiable function $\min_\beta f(\beta)$

$$\beta^{(k+1)} = \beta^{(k)} - \alpha_k \nabla f(\beta^{(k)})$$

The gradient step can be written equivalently as

$$\beta^{(k+1)} = \arg\min_\beta \{\underbrace{f(\beta^{(k)}) + \langle \nabla f(\beta^{(k)}), \beta - \beta^{(k)} \rangle}_{\text{linear approximation}} + \underbrace{\frac{1}{2\alpha_k}\|\beta - \beta^{(k)}\|^2}_{\text{proximal term}}\}$$



Optimizing composite functions

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad f(\beta) + g(\beta)$$

- $f : \mathbb{R}^p \to \mathbb{R}$ is convex and differentiable, ∇f is L-Lipschitz continuous
- $g : \mathbb{R}^p \to (-\infty, +\infty]$ is closed proper convex, non-differentiable
- For example, in lasso,

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \underbrace{\frac{1}{2}\|X\beta - Y\|^2}_{=f(\beta)} + \underbrace{\lambda\|\beta\|_1}_{=g(\beta)}$$

- Since g is non-differentiable, we cannot apply gradient methods

Proximal gradient step

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad f(\beta) + g(\beta)$$

Gradient step (suppose $g(\beta)$ disappears):

$$\beta^{(k+1)} = \beta^{(k)} - \alpha_k \nabla f(\beta^{(k)})$$

which can be written equivalently as

$$\beta^{(k+1)} = \arg\min_{\beta} \left\{ f(\beta^{(k)}) + \langle \nabla f(\beta^{(k)}), \beta - \beta^{(k)} \rangle + \frac{1}{2\alpha_k} \|\beta - \beta^{(k)}\|^2 \right\}$$

Proximal gradient step:

$$\beta^{(k+1)} = \arg\min_{\beta} \left\{ f(\beta^{(k)}) + \langle \nabla f(\beta^{(k)}), \beta - \beta^{(k)} \rangle + g(\beta) + \frac{1}{2\alpha_k} \|\beta - \beta^{(k)}\|^2 \right\}$$

Proximal gradient step

Proximal gradient step:

$$\beta^{(k+1)} = \arg\min_{\beta} \left\{ f(\beta^{(k)}) + \langle \nabla f(\beta^{(k)}), \beta - \beta^{(k)} \rangle + g(\beta) + \frac{1}{2\alpha_k} \|\beta - \beta^{(k)}\|^2 \right\}$$

After ignoring constant terms and completing the square, the above step can be written equivalently as

$$\begin{aligned} \beta^{(k+1)} &= \arg\min_{\beta} \left\{ \frac{1}{2\alpha_k} \left\| \beta - \left(\beta^{(k)} - \alpha_k \nabla f(\beta^{(k)}) \right) \right\|^2 + g(\beta) \right\} \\ &= P_{\alpha_k g} \left(\beta^{(k)} - \alpha_k \nabla f(\beta^{(k)}) \right) \end{aligned}$$

Proximal gradient step

Proximal gradient step:

$$\beta^{(k+1)} = \arg\min_{\beta} \left\{ f(\beta^{(k)}) + \langle \nabla f(\beta^{(k)}), \beta - \beta^{(k)} \rangle + g(\beta) + \frac{1}{2\alpha_k} \|\beta - \beta^{(k)}\|^2 \right\}$$

After ignoring constant terms and completing the square, the above step can be written equivalently as

$$\begin{aligned} \beta^{(k+1)} &= \arg\min_{\beta} \left\{ \frac{1}{2\alpha_k} \left\| \beta - \left(\beta^{(k)} - \alpha_k \nabla f(\beta^{(k)}) \right) \right\|^2 + g(\beta) \right\} \\ &= P_{\alpha_k g} \left(\beta^{(k)} - \alpha_k \nabla f(\beta^{(k)}) \right) \end{aligned}$$

Derivation* completing the square

$$\begin{aligned} &\langle \nabla f(\beta^{(k)}), \beta \rangle + \frac{1}{2\alpha_k} \|\beta - \beta^{(k)}\|^2 \\ =&\langle \nabla f(\beta^{(k)}) - \frac{1}{\alpha_k} \beta^{(k)}, \beta \rangle + \frac{1}{2\alpha_k} \|\beta\|^2 + \text{constant} \\ =&\frac{1}{\alpha_k} \langle \alpha_k \nabla f(\beta^{(k)}) - \beta^{(k)}, \beta \rangle + \frac{1}{2\alpha_k} \|\beta\|^2 + \text{constant} \end{aligned}$$

Proximal gradient methods

Algorithm (Proximal gradient (PG) method)

Choose $\beta^{(0)}$, constant step length $\alpha > 0$. Set $k \leftarrow 0$

repeat until convergence

$$\beta^{(k+1)} = P_{\alpha g}\left(\beta^{(k)} - \alpha \nabla f(\beta^{(k)})\right)$$

$$k \leftarrow k + 1$$

end(repeat)

return $\beta^{(k)}$

Proximal gradient methods

(Informal) in convex problems (f and g are convex), iteration complexity of PG method is $O(\frac{1}{k})$:

$$f(\beta^{(k)}) + g(\beta^{(k)}) - \underbrace{\min_{\beta \in \mathbb{R}^p} f(\beta) + g(\beta)}_{\text{optimal value}} \leq O\left(\frac{1}{k}\right)$$

If adopting stopping condition

$$f(\beta^{(k)}) + g(\beta^{(k)}) - \text{optimal value} \leq 10^{-4}$$

we need $O(10^4)$ PG iterations

Nesterov's accelerated method

Nesterov's idea: include a momentum term for acceleration

$$\bar{\beta}^{(k)} = \beta^{(k)} + \underbrace{\frac{t_k - 1}{t_{k+1}} \left(\beta^{(k)} - \beta^{(k-1)}\right)}_{\text{momentum term}}$$

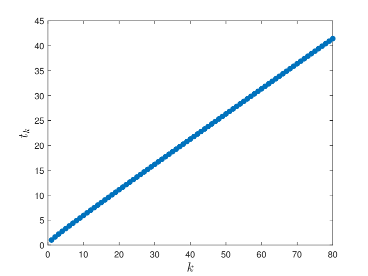
$$\beta^{(k+1)} = P_{\alpha g}\left(\bar{\beta}^{(k)} - \alpha \nabla f(\bar{\beta}^{(k)})\right)$$

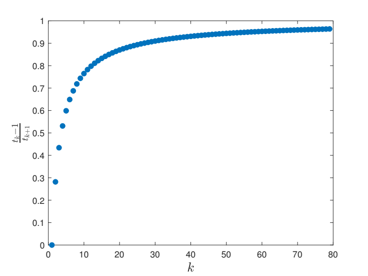
$\{t_k\}$ is a positive sequence such that $t_0 = t_1 = 1$, $t_{k+1}^2 - t_{k+1} \leq t_k$

- e.g., $t_k = 1, \forall\, k$. In this case, momentum term $= 0$, it is reduced to PG without acceleration
- e.g., $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2} \Rightarrow t_0 = 1, t_1 = 1, t_2 = \frac{1+\sqrt{5}}{2}, \ldots$
- there are many other sequences satisfying the condition
- Next we simply take $t_0 = t_1 = 1, t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$

Sequence t_k

Take $t_0 = t_1 = 1, t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$





Accelerated proximal gradient methods

Algorithm (Accelerated proximal gradient (APG) method)

Choose $\beta^{(0)}$, constant step length $\alpha > 0$. Set $t_0 = t_1 = 1$, $k \leftarrow 0$

repeat until convergence

$$\bar{\beta}^{(k)} = \beta^{(k)} + \frac{t_k - 1}{t_{k+1}}(\beta^{(k)} - \beta^{(k-1)})$$

$$\beta^{(k+1)} = P_{\alpha g}\left(\bar{\beta}^{(k)} - \alpha \nabla f(\bar{\beta}^{(k)})\right)$$

$$k \leftarrow k + 1$$

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

end(repeat)

return $\beta^{(k)}$

The algorithmic framework follows from [1]. It is built based on Nesterov's accelerated method in 1983 [2]

Accelerated proximal gradient methods

(Informal) in convex problems (f and g are convex), iteration complexity of APG method is $O(\frac{1}{k^2})$:

$$f(\beta^{(k)}) + g(\beta^{(k)}) - \underbrace{\min_{\beta \in \mathbb{R}^p} f(\beta) + g(\beta)}_{\text{optimal value}} \le O\left(\frac{1}{k^2}\right)$$

If adopting stopping condition

$$f(\beta^{(k)}) + g(\beta^{(k)}) - \text{optimal value} \le 10^{-4}$$

we need $O(10^2)$ PG iterations

Accelerated proximal gradient methods

- Backtracking line search is also applicable for finding step length α_k.
- For simplicity, we take a constant step length. f It should satisfy $\alpha \in (0, \frac{1}{L})$, where L is Lipschitz constant of $\nabla f(\cdot)$ (typically unknown)
- APG methods enjoy the same computational cost per iteration as PG methods.
- Iteration complexity: APG $O(\frac{1}{k^2})$; PG $O(\frac{1}{k})$

APG for lasso

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \underbrace{\frac{1}{2}\|X\beta - Y\|^2}_{=f(\beta)} + \underbrace{\lambda\|\beta\|_1}_{=g(\beta)} \quad \text{lasso problem}$$

Then $\nabla f(\beta) = X^T(X\beta - Y)$ with Lipschitz constant $L = \lambda_{\max}(X^T X)$. Choose step length $\alpha = 1/L$. APG iterations:

$$\bar{\beta}^{(k)} = \beta^{(k)} + \frac{t_k - 1}{t_{k+1}} \left(\beta^{(k)} - \beta^{(k-1)}\right)$$

$$\beta^{(k+1)} = S_{\lambda/L} \left(\bar{\beta}^{(k)} - \frac{1}{L} X^T (X\bar{\beta}^{(k)} - Y)\right)$$

APG is also applicable for "logistic regression + lasso regularization"

APG for lasso

Design a stopping criteria for lasso problem

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}} \quad \underbrace{\frac{1}{2}\|X\beta - Y\|^2}_{=f(\beta)} + \underbrace{\lambda\|\beta\|_1}_{=g(\beta)}$$

We know that β is a global minimizer to lasso problem if and only if

$$0 \in \nabla f(\beta) + \partial g(\beta),$$

which is equivalent to [check]

$$\beta = P_g(\beta - \nabla f(\beta)).$$

Therefore, we can choose a tolerance $\varepsilon > 0$ and stop the method at $\beta^{(k)}$ once the stopping criteria is satisfied

$$\left\|\beta^{(k)} - S_\lambda\left(\beta^{(k)} - X^T(X\beta^{(k)} - Y)\right)\right\| < \varepsilon.$$

Restart

- A strategy to speed up APG is to restart the algorithm after a fixed number of iterations
- using the latest iterate as the starting point of the new round of APG iteration
- a reasonable choice is to perform restart every 100 or 200 iterations

References i

A. Beck and M. Teboulle.
A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
SIAM journal on imaging sciences, 2(1):183–202, 2009.

Y. E. Nesterov.
A method for solving the convex programming problem with convergence rate O$(1/k^2)$.
In Dokl. Akad. Nauk SSSR,, volume 269, pages 543–547, 1983.