Introductory Course on Non-smooth Optimisation

Lecture 02 - Proximal gradient method

Outline

Outline

Problem

Problem (Unconstrained non-smooth optimisation)

Consider minising

$$\min_{x \in \mathbb{R}^n} R(x),$$

where $R : \mathbb{R}^n \to]-\infty, +\infty]$ *is proper convex and lower semi-continuous.*

Γ_0: the class of proper convex and lower semi-continuous functions on $\mathbb{R}^n$.

Problem

Problem (Unconstrained non-smooth optimisation)

Consider minising

$$\min_{x \in \mathbb{R}^n} R(x),$$

where $R : \mathbb{R}^n \to]-\infty, +\infty]$ *is proper convex and lower semi-continuous.*

Γ_0: the class of proper convex and lower semi-continuous functions on $\mathbb{R}^n$.

- The set of minimisers, *i.e.*

$$\mathrm{Argmin}(R) = \{x \in \mathbb{R}^n : R(x) = \min_{x \in \mathbb{R}^n} R(x)\},$$

  is non-empty

- $R(x)$ is non-differentiable...

Subdifferential

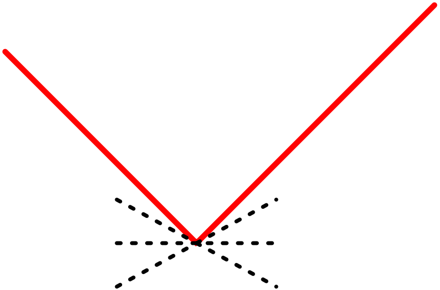
Definition

Let $R \in \Gamma_0$, the subdifferential of R at $x \in \mathrm{dom}(R)$ is defined by

$$\partial R : \mathbb{R}^n \rightrightarrows \mathbb{R}^n,\, x \mapsto \left\{ g \in \mathbb{R}^n \,|\, R(y) \geq R(x) + \langle g,\, y - x \rangle,\ \forall y \in \mathbb{R}^n \right\}.$$

Example:

$$\partial |x| = \begin{cases} +1 : x > 0 \\ [-1, 1] : x = 0 \\ -1 : x < 0 \end{cases}$$



Subdifferential

Definition

Let $R \in \Gamma_0$, the subdifferential of R at $x \in \mathrm{dom}(R)$ is defined by

$$\partial R : \mathbb{R}^n \rightrightarrows \mathbb{R}^n,\ x \mapsto \big\{ g \in \mathbb{R}^n \,|\, R(y) \geq R(x) + \langle g,\, y - x \rangle,\ \forall y \in \mathbb{R}^n \big\}.$$

Lemma

Let $R \in \Gamma_0$ and $x \in \mathrm{dom}(R)$, then

- $\partial R(x) = \big\{ g \in \mathbb{R}^n : R(y) \geq R(x) + \langle g,\, y - x \rangle \big\}$;
- $\partial R(x)$ *is closed and convex;*

Subdifferential

Definition

Let $R \in \Gamma_0$, the subdifferential of R at $x \in \mathrm{dom}(R)$ is defined by

$$\partial R : \mathbb{R}^n \rightrightarrows \mathbb{R}^n,\ x \mapsto \big\{ g \in \mathbb{R}^n \,|\, R(y) \geq R(x) + \langle g,\, y - x \rangle,\ \forall y \in \mathbb{R}^n \big\}.$$

Lemma

Let $R \in \Gamma_0$ and $x \in \mathrm{dom}(R)$, then

- $\partial R(x) = \big\{ g \in \mathbb{R}^n : R(y) \geq R(x) + \langle g,\, y - x \rangle \big\}$*;*
- $\partial R(x)$ *is closed and convex;*

Lemma (Monotonicity)

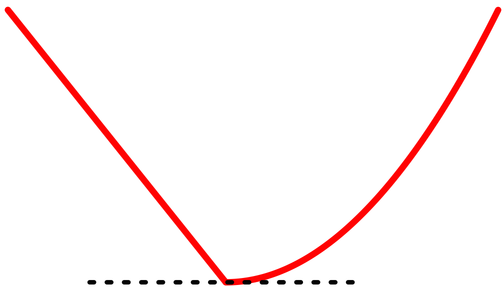
Let $R \in \Gamma_0$, then $\forall x, y \in \mathrm{dom}(R)$,

$$\langle u - v,\, x - y \rangle \geq 0,\ \forall u \in \partial R(x),\ v \in \partial R(y).$$

Optimality condition

$x^\star$ minimises $R(x)$ if and only if

$$0 \in \partial R(x^\star).$$



$$R(y) \geq R(x^\star) + \langle g,\, y - x \rangle \text{ holds for all } y \in \mathrm{dom}(R) \iff 0 \in \partial R(x^\star).$$

Subgradient descent

Subgradient descent

initial: $x_0 \in \mathrm{dom}(R)$;

repeat:

1. Choose step-size $\gamma_k > 0$ and a subgradient $g_k \in \partial R(x_k)$
2. Update $x_{k+1} = x_k - \gamma_k g_k$

until: stopping criterion is satisfied.

Subgradient descent

Subgradient descent

initial: $x_0 \in \mathrm{dom}(R)$;

repeat:

1. Choose step-size $\gamma_k > 0$ and a subgradient $g_k \in \partial R(x_k)$
2. Update $x_{k+1} = x_k - \gamma_k g_k$

until: stopping criterion is satisfied.

Step-size rule:

- Fixed step-size: γ_k is constant;
- Fixed length: $\gamma_k \|g_k\| = \|x_{k+1} - x_k\|$ is a constant;
- Diminishing step-size: $\gamma_k \to 0$, $\sum_i \gamma_i = +\infty$.

Assumptions

Assumptions:

- R has minimiser $x^\star$ and finite optimal value $R(x^\star)$;
- R is convex, $\mathrm{dom}(R) = \mathbb{R}^n$;
- R is Lipschitz consinuout with constant L:

$$|R(x) - R(y)| \leq L\|x - y\|, \ \forall x, y \in \mathrm{dom}(R). \tag{1.1}$$

Assumptions

Assumptions:

- R has minimiser $x^\star$ and finite optimal value $R(x^\star)$;
- R is convex, $\mathrm{dom}(R) = \mathbb{R}^n$;
- R is Lipschitz consinuout with constant L:

$$|R(x) - R(y)| \leq L\|x - y\|, \ \forall x, y \in \mathrm{dom}(R). \tag{1.1}$$

Eq. (1.1) implies $\|g\| \leq L$ for all $x \in \mathrm{dom}(R)$.

Convergence propergies

Subgradient descent is **NOT** a descent method.

Fixed step-size $\gamma_k \equiv \gamma$:

$$R_{k,best} - R(x^\star) \leq \frac{\|x_0 - x^\star\|^2}{2k\gamma} + \frac{\gamma L^2}{2}.$$

- Does not guarantee the convergence of $R_{k,best}$
- For large k, $R_{k,best}$ is approximately $\frac{\gamma L^2}{2}$ suboptimal

Convergence propergies

Subgradient descent is **NOT** a descent method.

Fixed step-length $\gamma_k = c/\|g_k\|$:

$$R_{k,best} - R(x^\star) \leq \frac{\|x_0 - x^\star\|^2}{2kc} + \frac{cL}{2}.$$

- Does not guarantee the convergence of $R_{k,best}$
- For large k, $R_{k,best}$ is approximately $\frac{cL}{2}$ suboptimal

Convergence propergies

Subgradient descent is **NOT** a descent method.

Diminishing step-size: $\gamma_k \to 0$, $\sum_i \gamma_i = +\infty$:

$$R_{k,best} - R(x^\star) \leq \frac{\|x_0 - x^\star\|^2 + L^2 \sum_{i=1}^k \gamma_i^2}{\sum_{i=1}^k \gamma_i}.$$

- If $\sum_{i=1}^k \gamma_i^2 / \sum_{i=1}^k \gamma_i \to 0$, then $R_{k,best} \to R(x^\star)$
- Choice of γ_k: $\gamma_k = c/k^q$, $q \in]1/2, 1[$

Optimal step-size

For fixed number of iterations: If $c_i = \gamma_i \|g_i\|$ and $\|x_0 - x^\star\| \leq D$,

$$R_{k,best} - R(x^\star) \leq \frac{D^2 + L^2 \sum_{i=1}^k c_i^2}{2 \sum_{i=1}^k \gamma_i / L}.$$

- For given k, rhs is minimised by $c_i = c = D/\sqrt{k}$
- Hence the rate

$$R_{k,best} - R(x^\star) \leq \frac{LD}{\sqrt{k}}.$$

- Iteration complexity: reach $R_{k,best} - R(x^\star) < \epsilon$ in $O(1/\epsilon^2)$ steps

Optimal step-size

For fixed number of iterations: If $c_i = \gamma_i \|g_i\|$ and $\|x_0 - x^\star\| \leq D$,

$$R_{k,best} - R(x^\star) \leq \frac{D^2 + L^2 \sum_{i=1}^k c_i^2}{2 \sum_{i=1}^k \gamma_i / L}.$$

- For given k, rhs is minimised by $c_i = c = D/\sqrt{k}$
- Hence the rate
$$R_{k,best} - R(x^\star) \leq \frac{LD}{\sqrt{k}}.$$
- Iteration complexity: reach $R_{k,best} - R(x^\star) < \epsilon$ in $O(1/\epsilon^2)$ steps

When $R(x^\star)$ is available: step-size

$$\gamma_k = \frac{R(x_k) - R(x^\star)}{\|g_k\|^2}.$$

Convergence rate:

$$R_{k,best} - R(x^\star) \leq \frac{LD}{\sqrt{k}}.$$

NB: $O(1/\sqrt{k})$ is the best rate can be obtained by subgradient method.

Remarks

- Handles non-smooth problem
- Simple iterative scheme
- Slow convergence rate
- No clear stopping criterion

NB: need a better approach to handle non-smoothness...

Outline

Projected gradient descent

Problem (Constrained smooth optimisation)

Let $F \in C_L^1$ and $\Omega \subseteq \mathbb{R}^n$ be a closed and convex set

$$\min_{x \in \Omega} F(x).$$

Projected gradient descent

Problem (Constrained smooth optimisation)

Let $F \in C_L^1$ and $\Omega \subseteq \mathbb{R}^n$ be a closed and convex set

$$\min_{x \in \Omega} F(x).$$

Projected gradient descent

initial: $x_0 \in \Omega$;

repeat:

1. Choose step-size $\gamma_k \in]0, 2/L[$
2. Gradient descent $x_{k+1/2} = x_k - \gamma_k \nabla F(x_k)$
3. Projection $x_{k+1} = \mathrm{proj}_\Omega(x_{k+1/2})$

until: stopping criterion is satisfied.

Composite optimisation problem

As $\iota_\Omega \in \Gamma_0$.

Problem (Composite optimisation)

Consider the following optimisation problem

$$\min_{x\in\mathbb{R}^n} \Big\{ \Phi(x) \stackrel{def}{=} R(x) + F(x) \Big\}.$$

Assumtions:

- $F \in C_L^1$
- $R \in \Gamma_0$
- $\mathrm{Argmin}(\Phi) \neq \emptyset$

Proximal gradient descent

Projection onto a set

$$\mathrm{proj}_\Omega(y) \stackrel{\text{def}}{=} \mathrm{argmin}_{x\in\mathbb{R}^n}\, \iota_\Omega(x) + \frac{1}{2}\|x-y\|^2.$$

Proximal mapping

$$\mathrm{proj}_{\gamma R}(y) \stackrel{\text{def}}{=} \mathrm{argmin}_{x\in\mathbb{R}^n}\, \gamma R(x) + \frac{1}{2}\|x-y\|^2.$$

Proximal gradient descent

Projection onto a set

$$\mathrm{proj}_{\Omega}(y) \stackrel{\text{def}}{=} \mathrm{argmin}_{x\in\mathbb{R}^n}\, \iota_{\Omega}(x) + \frac{1}{2}\|x-y\|^2.$$

Proximal mapping

$$\mathrm{proj}_{\gamma R}(y) \stackrel{\text{def}}{=} \mathrm{argmin}_{x\in\mathbb{R}^n}\, \gamma R(x) + \frac{1}{2}\|x-y\|^2.$$

Projected gradient descent

initial: $x_0 \in \Omega$;

repeat:

1. Choose step-size $\gamma_k \in]0, 2/L[$
2. Gradient descent $x_{k+1/2} = x_k - \gamma_k \nabla F(x_k)$
3. Projection $x_{k+1} = \mathrm{prox}_{\gamma_k R}(x_{k+1/2})$

until: stopping criterion is satisfied.

Interpretation

A.K.A Forward–Backward splitting:

- Forward step: gradient descent of F
- Backward step: proximity operator of R

Interpretation

A.K.A Forward–Backward splitting:

- Forward step: gradient descent of F
- Backward step: proximity operator of R

Interation in one line:

$$x_{k+1} = \mathrm{prox}_{\gamma_k R}(x_k - \gamma_k \nabla F(x_k)).$$

Definition of $\mathrm{prox}_{\gamma R}$,

$$\begin{aligned} x_{k+1} &= \mathrm{argmin}_x \left\{ \gamma_k R(x) + \tfrac{1}{2} \|x - (x_k - \gamma_k \nabla F(x_k))\|^2 \right\} \\ &= \mathrm{argmin}_x \left\{ \gamma_k R(x) + \gamma_k \langle \nabla F(x_k),\, x - x_k \rangle + \tfrac{1}{2} \|x - x_k\|^2 \right\} \\ &= \mathrm{argmin}_x \left\{ R(x) + \boxed{F(x_k) + \langle \nabla F(x_k),\, x - x_k \rangle + \frac{1}{2\gamma_k} \|x - x_k\|^2} \right\} \end{aligned}$$

NB: x_{k+1} minimises $R(x)$ plus the majorisation of $F(x)$ at x_k

Special cases

Gradient descent: $R = 0$

$$x_{k+1} = x_k - \gamma_k \nabla F(x_k).$$

Proximal point algorithm: $F = 0$

$$x_{k+1} = \mathrm{prox}_{\gamma_k R}(x_k).$$

Projected gradient descent: $R = \iota_\Omega$

$$x_{k+1} = \mathrm{proj}_\Omega(x_k - \gamma_k \nabla F(x_k)).$$

ISTA: iterative shrinkage-thresholding algorithm: $R = \|x\|_1$

$$x_{k+1} = \mathcal{T}_\gamma(x_k - \gamma_k \nabla F(x_k)),$$

where

$$\left(\mathrm{prox}_{\gamma\|\cdot\|_1}(y)\right)_i = \begin{cases} \mathrm{sign}(y_i) \cdot (|y_i| - \gamma) : |y_i| > \gamma \\ \hfill 0 : y_i \in [-\gamma, \gamma]. \end{cases}$$

Two basic lemmas

Define

$$E_\gamma(x, y) \stackrel{\text{def}}{=} R(x) + F(y) + \langle \nabla F(y),\, x - y \rangle + \frac{1}{2\gamma}\|x - y\|^2$$

and $y_+ \stackrel{\text{def}}{=} \mathrm{argmin}_x E_\gamma(x, y)$.

Lemma

Let $y \in \mathbb{R}^n$ and $\gamma \in]0, 2/L[$ such that

$$\Phi(y_+) \leq E_\gamma(y_+, y),$$

then for any $x \in \mathbb{R}^n$,

$$\Phi(x) - \Phi(y_+) \geq \frac{1}{2\gamma}\big(\|x - y_+\|^2 - \|x - y\|^2\big).$$

Lemma

Given $y \in \mathbb{R}^n$ and $\gamma \in]0, 1/L]$, then for any $x \in \mathbb{R}^n$,

$$\Phi(y_+) + \frac{1}{2\gamma}\|y_+ - x\|^2 \leq \Phi(y) + \frac{1}{2\gamma}\|y - x\|^2.$$

Convergence analysis

Consequence: proximal gradient is a descent method.

Consider $\gamma_k \equiv \gamma \in]0, 1/L]$

- For each step

$$\Phi(x_k) - \Phi(x_{k+1}) \geq \frac{\gamma}{2}\|x_k - x_{k+1}\|^2.$$

- Regarding $\Phi(x^\star)$

$$\Phi(x_{k+1}) - \Phi(x^\star) \leq \frac{\gamma}{2}\big(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2\big).$$

- Summing up

$$\begin{aligned} k\big(\Phi(x_k) - \Phi(x^\star)\big) &\leq \sum_{i=1}^{k}\big(\Phi(x_i) - \Phi(x^\star)\big) \\ &\leq \frac{\gamma}{2}\sum_{i=1}^{k}\big(\|x_{i-1} - x^\star\|^2 - \|x_i - x^\star\|^2\big) \leq \frac{\gamma}{2}\|x_0 - x^\star\|^2 \end{aligned}$$

- $O(1/k)$ rate

$$\Phi(x_k) - \Phi(x^\star) \leq \frac{\gamma\|x_0 - x^\star\|^2}{2k}.$$

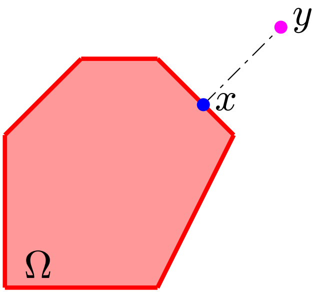
NB: not optimal and can be accelerated

Outline

Projection onto sets

Indicator function: let $\Omega \subseteq \mathbb{R}^n$

$$\iota_\Omega(x) = \begin{cases} 0 : x \in \Omega, \\ +\infty : x \notin \Omega. \end{cases}$$



Projection of y onto Ω:

$$\min_{x \in \Omega} \|x - y\|.$$

Definition (Projection)

Projection mapping onto a set is defined by

$$\mathrm{proj}_\Omega(y) \stackrel{\mathrm{def}}{=} \mathrm{argmin}_{x \in \mathbb{R}^n} \iota_\Omega(x) + \|x - y\|^2.$$

From projection to proximal mapping

Definition (Proximal mapping)

The proimal mapping (proximity operator) of a fucntion $R \in \Gamma_0$ is defined by

$$\mathrm{prox}_{\gamma R}(y) \stackrel{\text{def}}{=} \mathrm{argmin}_{x \in \mathbb{R}^n} \gamma R(x) + \frac{1}{2}\|x - y\|^2.$$

Optimality condition: denote $y_+ \stackrel{\text{def}}{=} \mathrm{prox}_R(y)$,

$$0 \in \partial R(y_+) + y_+ - y \iff y_+ = (\mathrm{Id} + \partial R)^{-1}(y).$$

From projection to proximal mapping

Definition (Proximal mapping)

The proimal mapping (proximity operator) of a fucntion $R \in \Gamma_0$ is defined by

$$\mathrm{prox}_{\gamma R}(y) \stackrel{\text{def}}{=} \mathrm{argmin}_{x \in \mathbb{R}^n} \gamma R(x) + \frac{1}{2}\|x - y\|^2.$$

Optimality condition: denote $y_+ \stackrel{\text{def}}{=} \mathrm{prox}_R(y)$,

$$0 \in \partial R(y_+) + y_+ - y \iff y_+ = (\mathrm{Id} + \partial R)^{-1}(y).$$

$\mathcal{J}_{\partial R} \stackrel{\text{def}}{=} (\mathrm{Id} + \partial R)^{-1}$ is called the resolvent of ∂R:

- $\mathcal{J}_{\partial R}$ is firmly non-expansive

$$\langle \mathcal{J}_{\partial R}(x) - \mathcal{J}_{\partial R}(y), x - y \rangle \geq \|\mathcal{J}_{\partial R}(x) - \mathcal{J}_{\partial R}(y)\|^2.$$

- Reflection $\mathcal{R}_{\partial R} \stackrel{\text{def}}{=} 2\mathcal{J}_{\partial R} - \mathrm{Id}$ is non-expansive, *i.e.* 1-Lipschitz.

Examples

Projection: $R(x) = \iota_\Omega(x)$,

$$\mathrm{prox}_{\gamma R} = (\mathrm{Id} + \mathcal{N}_\Omega)^{-1} = \mathrm{proj}_\Omega.$$

Simple instances:

- Hyperplane: $\Omega = \{x : a^T x = b\}$, $a \neq 0$

$$\mathrm{proj}_\Omega = x + \frac{b - a^T x}{\|a\|^2} a.$$

- Affine subspace: $\Omega = \{x : Ax = b\}$ with $A \in \mathbb{R}^{m \times n}$, $\mathrm{rank}(A) = m < n$

$$\mathrm{proj}_\Omega = x + A^T (AA^T)^{-1} (b - Ax).$$

- Half space: $\Omega = \{x : a^T x \leq b\}$, $a \neq 0$

$$\mathrm{proj}_\Omega = x + \frac{b - a^T x}{\|a\|^2} a \ \text{ if } a^T x > b \quad \text{and} \quad x \ \text{ if } a^T x \leq b.$$

- Nonnegative orthant: $\Omega = \mathbb{R}^n_+$

$$\mathrm{proj}_\Omega = \big(\max\{0, x_i\}\big)_i.$$

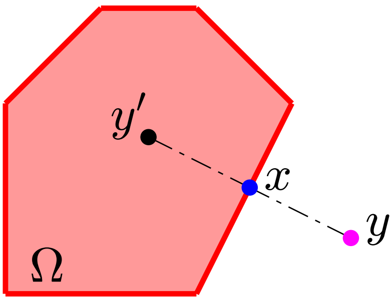
Examples

Projection: $R(x) = \iota_\Omega(x)$,

$$\mathrm{prox}_{\gamma R} = (\mathrm{Id} + \mathcal{N}_\Omega)^{-1} = \mathrm{proj}_\Omega.$$

Reflection

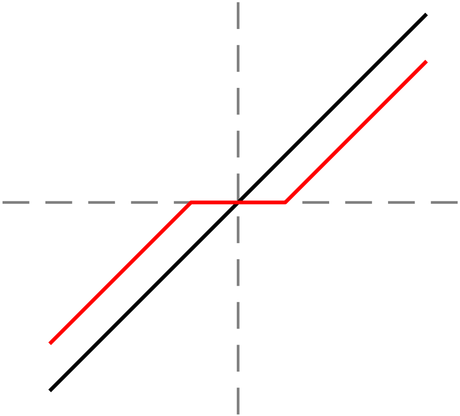
$$\mathcal{R}_{\mathcal{N}_\Omega} = 2\mathrm{proj}_\Omega - \mathrm{Id} = \mathrm{proj}_\Omega + (\mathrm{proj}_\Omega - \mathrm{Id}).$$



Examples

Soft-threshold: $R(x) = |x|$,

$$\operatorname{prox}_{\gamma R}(y) = \begin{cases} y - \gamma : y > \gamma \\ 0 : y \in [-\gamma, \gamma] \\ y + \gamma : y < -\gamma. \end{cases}$$



Examples

Quadratic function: $R(x) = \frac{1}{2}x^T Ax + b^T x + c,\ A \succeq 0$

$$\mathrm{prox}_{\gamma R}(y) = (\mathrm{Id} + \gamma A)^{-1}(y - \gamma b).$$

Euclidean norm: $R(x) = \|x\|$

$$\mathrm{prox}_{\gamma R}(y) = \begin{cases} (1 - \frac{\gamma}{\|y\|})y : \|y\| > \gamma \\ 0 : o.w. \end{cases}.$$

Logarithmic barrier: $R(x) = -\sum_i \log(x_i)$

$$\big(\mathrm{prox}_{\gamma R}(y)\big)_i = \frac{y_i + \sqrt{y_i^2 + 4\gamma}}{2},\ \ i = 1, ..., n.$$

Nuclear norm: $R(x) = \sum_i \delta_i$

$$\mathrm{prox}_{\gamma R}(y) = U\mathrm{prox}_{\gamma\|\cdot\|_1}\big(\mathrm{diag}(\Sigma)\big)V^T.$$

Calculus rules

Quadratic perturbation: $H(x) = R(x) + \frac{\alpha}{2}\|x\|^2 + \langle x,\ u\rangle + \beta$, $\alpha \geq 0$

$$\mathrm{prox}_H = \mathrm{prox}_{R/(\alpha+1)}\Big(\frac{x-u}{\alpha+1}\Big).$$

Translation: $H(x) = R(x - z)$

$$\mathrm{prox}_H = z + \mathrm{prox}_R(x - z).$$

Scaling: $H(x) = R(x/\rho)$

$$\mathrm{prox}_H = \rho\mathrm{prox}_{R/\rho^2}\Big(\frac{x}{\rho}\Big).$$

Reflection: $H(x) = R(-x)$

$$\mathrm{prox}_H = -\mathrm{prox}_R(-x).$$

Composition: $H(x) = R \circ L$ with L being bijective bounded linear mapping such that $L^{-1} = L^*$,

$$\mathrm{prox}_H = L^* \circ \mathrm{prox}_R \circ L.$$

Outline

From heavy-ball to inertial proximal gradient

An inertial proximal gradient

Initial: $x_0 \in \mathbb{R}^n$ and $\gamma \in]0, 2/L[$;

$$y_k = x_k + a_k(x_k - x_{k-1}),\ a_k \in [0, 1],$$
$$x_{k+1} = \mathrm{prox}_{\gamma R}(y_k - \gamma \nabla F(x_k)).$$

- Recovers inertial proximal point algorithm when $F = 0$, and heavy-ball method when $R = 0$
- Convergence via studying the Lyapunov function
$$\mathcal{E}(x_k) \stackrel{\text{def}}{=} \Phi(x_k) + \frac{a_k}{2\gamma}\|x_k - x_{k-1}\|^2.$$
- In general, no convergence rate.

A general inertial scheme

A general inertial proximal gradient

Initial: $x_0 \in \mathbb{R}^n, x_{-1} = x_0$ and $\gamma \in]0, 2/L[$;

$$y_k = x_k + a_k(x_k - x_{k-1}),\ a_k \in [0,1],$$
$$z_k = x_k + b_k(x_k - x_{k-1}),\ b_k \in [0,1],$$
$$x_{k+1} = \mathrm{prox}_{\gamma R}(y_k - \gamma \nabla F(z_k)).$$

- Convergence via studying the Lyapunov function

$$\mathcal{E}(x_k) \stackrel{\text{def}}{=} \Phi(x_k) + \tfrac{a_k}{2\gamma}\|x_k - x_{k-1}\|^2.$$

- In general, no convergence rate.
- Can be extend to multi-step, *e.g.*

$$y_k = x_k + a_{0,k}(x_k - x_{k-1}) + a_{1,k}(x_{k-1} - x_{k-2}) + \cdots .$$

Convergence rate

Assumption: $R = 0$, $F = \frac{1}{2}\|Ax - f\|^2$ and $(a_k, b_k) \equiv (a, b)$.

- $A^T A$ is symmetric positive definite with $A^T A \succeq \alpha \mathrm{Id}$;
- Taylor expansion
$$x_{k+1} = y_k - \gamma \nabla^2 F(x^\star)(z_k - x^\star).$$
- Let $d_k = (x_k - x^\star, x_{k-1} - x^\star)^T$ and $H = \nabla^2 F$, $G = \mathrm{Id} - \gamma H$, then
$$d_{k+1} = \underbrace{\begin{bmatrix} (a-b)\mathrm{Id} + (1+b)G, & -(a-b)\mathrm{Id} - bG \\ \mathrm{Id}, & 0 \end{bmatrix}}_{M} d_k.$$
- Spectral radius: $\eta = \rho(G) = 1 - \gamma\alpha$ and $\rho = \rho(M)$...

Spectral analysis

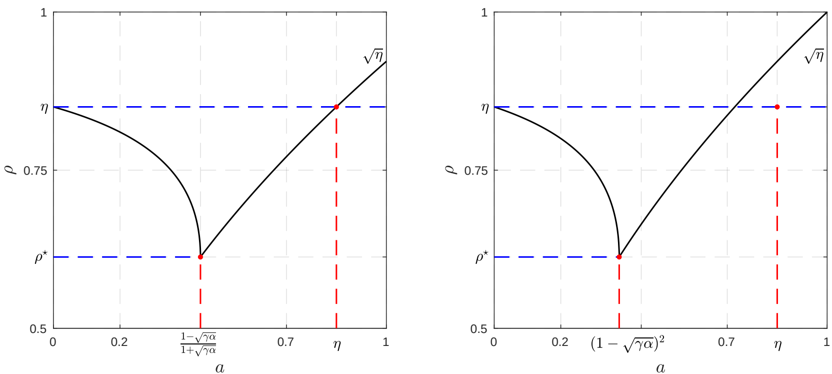
Lemma (Spectral radius ρ)

Between η and ρ,

- *η and ρ satisfy the relation*

$$\rho^2 - \big((a-b)+(1+b)\eta\big)\rho + (a-b) + b\eta = 0.$$

- *Given any $(a,b) \in [0,1[^2$, then $\rho(M) < 1$ if, and only if $\frac{2(b-a)-1}{1+2b} < \eta$.*



$b = a$ $\qquad\qquad$ $b = 0$

Remarks

- Given $b \in [0, 1]$, there exists optimal choice of $a \in [0, 1]$ such that

$$\rho = 1 - \sqrt{\gamma\alpha}$$

can be obtained.

- Take $b = a$, for

$$a \in \left]\tfrac{1-\sqrt{\gamma\alpha}}{1+\sqrt{\gamma\alpha}}, 1\right],$$

the leading eigenvalue of M is complex.

- Continue $b = a$, for

$$a \in]\eta, 1],$$

the inertial scheme is actually slower than the original scheme.

Outline

FISTA: fast iterative shrinkage-thresholding algorith

FISTA

Initial: $x_0 \in \mathbb{R}^n, x_{-1} = x_0$, $\gamma = 1/L$ and $t_0 = 1$;

$$t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2}, \;\; a_k = \frac{t_{k-1}-1}{t_k},$$

$$y_k = x_k + a_k(x_k - x_{k-1}),$$

$$x_{k+1} = \mathrm{prox}_{\gamma R}(y_k - \gamma \nabla F(y_k)).$$

- $t_k \approx \frac{k+1}{2}$
- $a_k \to 1$

Relation with Nesterov's optimal scheme

Nesterov: compute $\phi_k \in]0,1[$ from equation

$$\phi_k^2 = (1-\phi_k)\phi_{k-1}^2$$

and $a_k = \frac{\phi_{k-1}(1-\phi_{k-1})}{\phi_{k-1}^2+\phi_k}$.

- ϕ_k reads

$$\phi_k = \frac{-\phi_{k-1}^2 + \sqrt{\phi_{k-1}^4 + 4\phi_{k-1}^2}}{2} = \frac{2\phi_{k-1}^2}{\phi_{k-1}^2 + \sqrt{\phi_{k-1}^4 + 4\phi_{k-1}^2}}.$$

- Let $t_k = 1/\phi_k$,

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

- Directly

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

- And $a_k = \dfrac{t_{k-1}-1}{t_k}$.

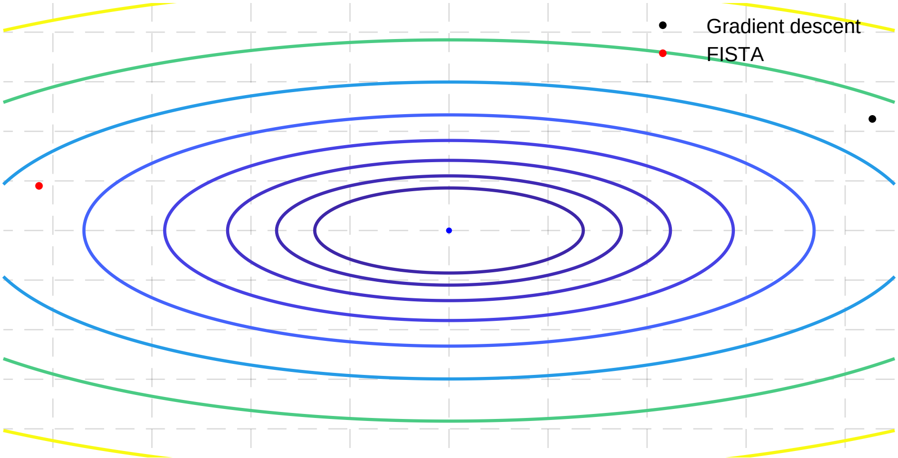
Convergence rate

NB: FISTA is not a descent method.

- Denote $f_k = \Phi(x_k) - \Phi(x^\star)$ and $u_k = t_k x_k - (t_k - 1)x_{k-1} - x^\star$, then
$$\frac{2}{L} t_k^2 f_k - \frac{2}{L} t_{k+1}^2 f_{k+1} \geq \|u_{k+1}\|^2 - \|u_k\|^2.$$
- Let c_k, d_k be positive sequences, if
$$c_k - c_{k+1} \geq d_{k+1} - d_k \forall k \geq 1, \text{ with} c_1 + d_1 < C, \ C > 0$$
then $c_k < C$ for all $k \geq 1$.
- $\frac{2}{L} t_k^2 f_k \leq \|x_0 - x^\star\|^2$
- $t_k \geq \frac{k+1}{2}$,
$$\Phi(x_k) - \Phi(x^\star) \leq \frac{L\|x_0 - x^\star\|^2}{2(k+1)^2}.$$

Outline

Oscillation of FISTA



Restarting FISTA

Why:

- for LSE, leading eigenvalue fo the system si complex.
- over extropolation, momentum beats gradient.

Restarting FISTA

Initial: $x_0 \in \mathbb{R}^n, x_{-1} = x_0$, $\gamma = 1/L$ and $t_0 = 1$;

repeat:

1. Run FISTA iteration
2. If $\langle y_k - x_{k+1},\, x_{k+1} - x_k \rangle > 0$: $t_k = 1, y_k = x_k$.

until: stopping criterion is satisfied.

Outline

ℓ_1-regularised least square

$$\min_{x\in\mathbb{R}^n} \mu\|x\|_1 + \frac{1}{2}\|\mathcal{K}x - f\|^2.$$

Sparse logistic regression

$$\min_{x\in\mathbb{R}^n} \mu\|x\|_1 + \frac{1}{m}\sum_{i=1}^m \log\left(1 + e^{-l_i h_i^T x}\right),$$

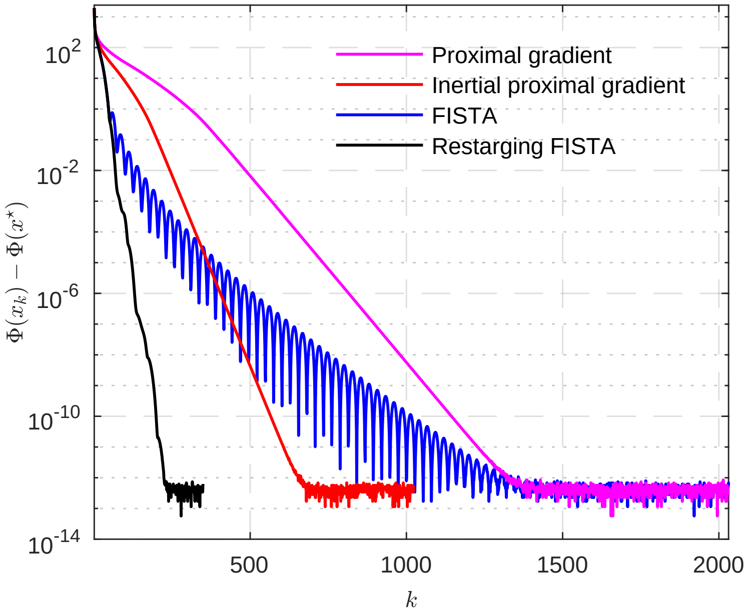
where $\mu = 10^{-2}$. The `australian` data set from LIBSVM[1] is considered.

[1] `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`

Compared methods

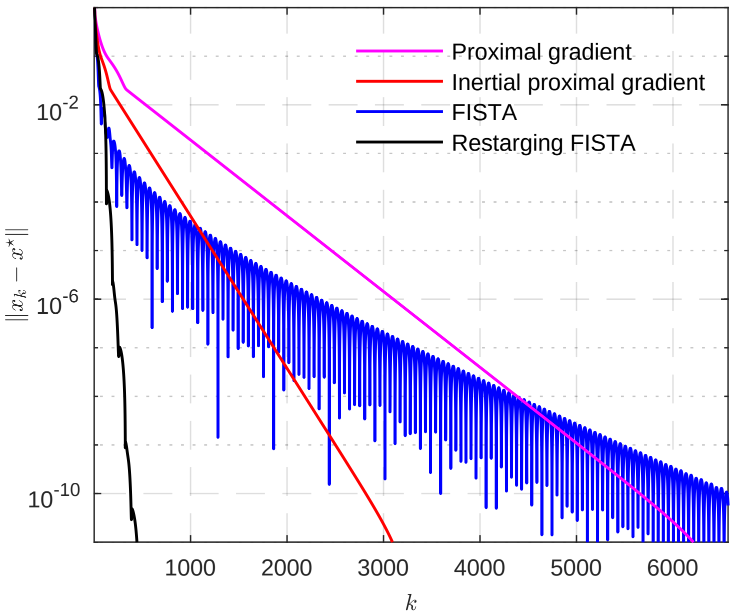
- Proximal gradient descent
- Inertial proximal gradient descent
- FISTA
- Restarting FISTA

Numerical results



ℓ_1-regularised LSE

Numerical results



Sparse logistic regression

Reference

- B. Polyak. Introduction to optimization. Optimization Software, 1987.
- Y. Nesterov. Introductory lectures on convex optimization: A basic course. Vol. 87. Springer Science & Business Media, 2013.
- 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.
- H. Bauschke and P. L. Combettes. Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, 2011.
- B. O'Donoghue and E. J. Candés. Adaptive restart for accelerated gradient schemes. Foundations of Computational Mathematics, pages 1–18, 2012.