Convex Optimization and Applications

12 - First Order Methods

Guillaume Sagnol

Technische Universität Berlin

Nonsmooth Convex Optimization

Many convex problems that arise in machine learning / signal processing are

- Non-smooth
- Unconstrained (or constrained over a very simple set)
- When the dimension of the problem is very large, the Newton steps of interior point methods become too expensive in practice.
- $\rightarrow$ preference given to first-order algorithms, that quickly converge to a reasonably good solution.

Examples (1/2)

- Lasso regression

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\textbf{minimize}}\ \|X\boldsymbol{\theta}-\boldsymbol{y}\|^2+\lambda\|\boldsymbol{\theta}\|_1,$$

- Soft-margin SVM

$$\underset{\boldsymbol{w}\in\mathbb{R}^n, b\in\mathbb{R}}{\textbf{minimize}}\quad \sum_{i=1}^m \max(0, 1-y_i(\boldsymbol{w}^T\boldsymbol{x}_i-b))+\lambda\|\boldsymbol{w}\|^2.$$

- D-optimal design

$$\begin{aligned}\underset{\boldsymbol{w}\in\mathbb{R}^n}{\textbf{maximize}}\quad & \det{}^{\frac{1}{n}}\Big(\sum_{i=1}^m w_i\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)\\ s.t.\quad & \boldsymbol{w}\geq\boldsymbol{0},\ \sum_i w_i=1.\end{aligned}$$

Examples (2/2)

- Low-rank Matrix completion

$$\underset{Y\in\mathbb{R}^{m\times n}}{\textbf{minimize}} \quad \sum_{(i,j)\in\Omega} (X_{ij} - Y_{ij})^2 + \lambda \|Y\|_*,$$

  The nuclear-norm $\|Y\|_* := \textbf{trace}\,(Y^T Y)^{\frac{1}{2}}$ serves as a convex approximation for **rank** Y.

- Total-Variation denoising

$$\underset{Y\in\mathbb{R}^{m\times n}}{\textbf{minimize}} \quad \|X - Y\|_F^2 + \lambda\, \mathrm{TV}(Y),$$

  The total-variation $\mathrm{TV}(Y) := \sum_{i,j} \left\| \begin{bmatrix} y_{i+1,j} - y_{i,j} \\ y_{i,j+1} - y_{i,j} \end{bmatrix} \right\|_2$ penalizes the pixels with a high local variation (*noise*).

Outline

Gradient descent

- A first order method is an algorithm to minimize a function F, that only uses first-oder derivatives
- The typical algorithm is *the gradient descent* [Cauchy, mid-19th]

$$x^{(k)} = x^{(k-1)} - t_k \nabla F(x^{(k-1)}),$$

  where the stepsize t_k is selected with a line search procedure.
- Obviously, we cannot use this method for *Nonsmooth optimization*.
- Hence, the most natural idea is to use *subgradients* instead.

Subgradient

Definition (Subgradient).

The vector $\boldsymbol{g}$ is a *subgradient* of $f : \mathbb{R}^n \to \mathbb{R}$ at $\boldsymbol{x} \in \mathbf{dom}\, f$, if

$$f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \boldsymbol{g}^T(\boldsymbol{z} - \boldsymbol{x}), \quad \forall \boldsymbol{z} \in \mathbf{dom}\, f.$$

Geometrically, this means that the vector $[\boldsymbol{g}, -1]^T$ defines a supporting hyperplane to **epi** f at $(\boldsymbol{x}, f(\boldsymbol{x}))$.

The *subdifferential* of f at $\boldsymbol{x}$ is the set of all subgradients:

$$\partial f(\boldsymbol{x}) := \{\boldsymbol{g} \in \mathbb{R}^n : \boldsymbol{g} \text{ is a subgradient of } f \text{ at } \boldsymbol{x}\}.$$

Properties

Proposition

(i) $\partial f(x)$ is always a closed convex set.

(ii) f convex $\implies \partial f(x) \neq \emptyset, \forall x \in \textbf{int dom} f$.

(iii) Let f be convex and $x \in \textbf{int dom} f$. Then,

$$f \text{ differentiable at } x \iff \partial f(x) = \{\nabla f(x)\}.$$

- For (i), we use that

$$\partial f(x) = \bigcap_{z \in \textbf{dom} f} \{\boldsymbol{g} : f(z) \geq f(x) + \boldsymbol{g}^T(z - x)\}$$

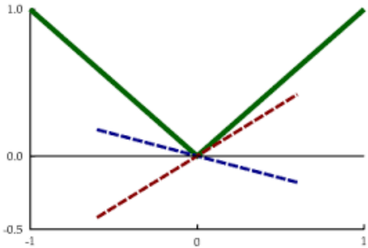
  is an intersection of halfspaces.

- (ii) follows from the supporting hyperplane theorem.

Example

Let $f : \mathbb{R} \to \mathbb{R}, x \mapsto |x|$. Then, the subdifferential of f is given by

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



Subgradient and optimality

Theorem

Let f be convex. Then, x^* minimizes f over $\mathbb{R}^n$ if and only if $\mathbf{0} \in \partial f(x^*)$.

Proof:

$$\mathbf{0} \in \partial f(x^*) \iff f(z) \geq f(x^*) + \underbrace{\mathbf{0}^T(z - x^*)}_{=0}, \ \forall z \in \mathbf{dom}\, f.$$

Calculus rules for subdifferentials

Let $f, f_1, \ldots, f_m$ be **convex**.

- Nonnegative scaling:

$$\partial(\alpha f)(\boldsymbol{x}) = \alpha \partial f(\boldsymbol{x}), \quad \text{for all } \alpha \geq 0.$$

Calculus rules for subdifferentials

Let $f, f_1, \ldots, f_m$ be **convex**.

- Nonnegative scaling:

$$\partial(\alpha f)(\boldsymbol{x}) = \alpha \partial f(\boldsymbol{x}), \quad \text{for all} \quad \alpha \geq 0.$$

- Sum:

$$\partial(f_1 + \ldots + f_m)(\boldsymbol{x}) = \partial f_1(\boldsymbol{x}) + \ldots + \partial f_m(\boldsymbol{x}).$$

(Note: this is a Minkowski sum of convex sets).

Calculus rules for subdifferentials

Let $f, f_1, \ldots, f_m$ be **convex.**

- Nonnegative scaling:
$$\partial(\alpha f)(x) = \alpha \partial f(x), \quad \text{for all } \alpha \geq 0.$$
- Sum:
$$\partial(f_1 + \ldots + f_m)(x) = \partial f_1(x) + \ldots + \partial f_m(x).$$
(Note: this is a Minkowski sum of convex sets).
- Affine transformation:
$$\partial\big(z \mapsto f(Az + b)\big)(x) = A^T \partial f(Ax + b).$$

Calculus rules for subdifferentials

Let $f, f_1, \ldots, f_m$ be **convex**.

- Nonnegative scaling:
$$\partial(\alpha f)(\boldsymbol{x}) = \alpha \partial f(\boldsymbol{x}), \quad \text{for all } \alpha \geq 0.$$
- Sum:
$$\partial(f_1 + \ldots + f_m)(\boldsymbol{x}) = \partial f_1(\boldsymbol{x}) + \ldots + \partial f_m(\boldsymbol{x}).$$
(Note: this is a Minkowski sum of convex sets).
- Affine transformation:
$$\partial\big(z \mapsto f(Az+b)\big)(\boldsymbol{x}) = A^T \partial f(A\boldsymbol{x}+b).$$
- Pointwise maximum:
Let $g(\boldsymbol{x}) = \max_{i=1,\ldots,m} f_i(\boldsymbol{x})$. Then, $\partial g(\boldsymbol{x}) = \textbf{conv}\,\Big(\bigcup_{j \in A(\boldsymbol{x})} \partial f_j(\boldsymbol{x})\Big)$,
where $A(\boldsymbol{x})$ is the set of *active functions* at $\boldsymbol{x}$, i.e.,
$A(\boldsymbol{x}) := \{j \in [m] : f_j(\boldsymbol{x}) = g(\boldsymbol{x})\}$.

(can be extended to pointwise supremums of infinitely many functions under additional technical conditions).

The subgradient method

To minimize a non-smooth convex function F over $\mathbb{R}^n$, we can use the subgradient method:

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} - t_k \boldsymbol{g}^{(k-1)}, \qquad \text{for some } \boldsymbol{g}^{(k-1)} \in \partial F(\boldsymbol{x}^{(k-1)}).$$

A few properties of this algorithm:

- Not a *descent method* (we can have $F(\boldsymbol{x}^{(k)}) > F(\boldsymbol{x}^{(k-1)})$, even for arbitrarily small step sizes.)

The subgradient method

To minimize a non-smooth convex function F over $\mathbb{R}^n$, we can use the subgradient method:

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} - t_k \boldsymbol{g}^{(k-1)}, \qquad \text{for some } \boldsymbol{g}^{(k-1)} \in \partial F(\boldsymbol{x}^{(k-1)}).$$

A few properties of this algorithm:

- Not a *descent method* (we can have $F(\boldsymbol{x}^{(k)}) > F(\boldsymbol{x}^{(k-1)})$, even for arbitrarily small step sizes.)
- Method can fail to converge if we use exact or backtracking line search.

The subgradient method

To minimize a non-smooth convex function F over $\mathbb{R}^n$, we can use the subgradient method:

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} - t_k \boldsymbol{g}^{(k-1)}, \qquad \text{for some } \boldsymbol{g}^{(k-1)} \in \partial F(\boldsymbol{x}^{(k-1)}).$$

A few properties of this algorithm:

- Not a *descent method* (we can have $F(\boldsymbol{x}^{(k)}) > F(\boldsymbol{x}^{(k-1)})$, even for arbitrarily small step sizes.)
- Method can fail to converge if we use exact or backtracking line search.
- Convergence can be proved for some *offline rules*, e.g.
  - Constant step sizes ($t_k = t > 0, \forall k \in \mathbb{N}$).
  - Nonsummable diminishing ($t_k \to 0, \sum_{k \in \mathbb{N}} t_k = \infty$)

The subgradient method

To minimize a non-smooth convex function F over $\mathbb{R}^n$, we can use the subgradient method:

$$\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} - t_k \boldsymbol{g}^{(k-1)}, \qquad \text{for some } \boldsymbol{g}^{(k-1)} \in \partial F(\boldsymbol{x}^{(k-1)}).$$

A few properties of this algorithm:

- Not a *descent method* (we can have $F(\boldsymbol{x}^{(k)}) > F(\boldsymbol{x}^{(k-1)})$, even for arbitrarily small step sizes.)
- Method can fail to converge if we use exact or backtracking line search.
- Convergence can be proved for some *offline rules*, e.g.
  - Constant step sizes ($t_k = t > 0, \forall k \in \mathbb{N}$).
  - Nonsummable diminishing ($t_k \to 0, \sum_{k\in\mathbb{N}} t_k = \infty$)
- Convergence typically *slow*: after k iteration, the best iterate seen so far satisfies $f(\boldsymbol{x}^{(k)}_{\mathrm{best}}) \leq f(\boldsymbol{x}^*) + O(\frac{1}{\sqrt{k}})$.

Outline

Strong convexity

Definition (ν-strong convexity).

f is ν-strongly convex for some $\nu > 0$ iff $x \mapsto f(x) - \frac{\nu}{2}\|x\|^2$ is convex.

Remark: If f is twice diff., then f is ν-strongly convex iff

$$\nabla^2 f(x) \succeq \nu I, \ \forall x \in \mathbf{dom}\, f.$$

Strong convexity

Definition (ν-strong convexity).

f is ν-strongly convex for some $\nu > 0$ iff $x \mapsto f(x) - \frac{\nu}{2}\|x\|^2$ is convex.

Remark: If f is twice diff., then f is ν-strongly convex iff

$$\nabla^2 f(x) \succeq \nu I, \ \forall x \in \mathbf{dom}\, f.$$

Proposition

Let f be ν-strongly convex. Then, $\forall x_0 \in \mathbf{dom}\, \partial f, \forall \boldsymbol{g} \in \partial f(x_0)$,

$$f(x) \geq f(x_0) + \langle \boldsymbol{g}, x - x_0 \rangle + \frac{\nu}{2}\|x - x_0\|^2, \ \forall x \in \mathbf{dom}\, f.$$

Remark: The converse statement is also true.

Strong convexity

Proposition

Let f be ν-strongly convex. Then, $\forall \boldsymbol{x}_0 \in \mathbf{dom}\, \partial f$, $\forall \boldsymbol{g} \in \partial f(\boldsymbol{x}_0)$,

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0) + \langle \boldsymbol{g}, \boldsymbol{x} - \boldsymbol{x}_0 \rangle + \frac{\nu}{2} \|\boldsymbol{x} - \boldsymbol{x}_0\|^2, \ \forall \boldsymbol{x} \in \mathbf{dom}\, f.$$

Proof:

- Let $F(\boldsymbol{x}) := f(\boldsymbol{x}) - \frac{\nu}{2}\|\boldsymbol{x}\|^2$; this is a convex function.
- Rule for sum of subdifferentials of convex functions:

$$\boldsymbol{g} \in \partial f(\boldsymbol{x}_0) = \partial F(\boldsymbol{x}_0) + \frac{\nu}{2} \nabla(\boldsymbol{x} \mapsto \|\boldsymbol{x}\|^2) = \partial F(\boldsymbol{x}_0) + \nu \boldsymbol{x}_0$$

- So $\boldsymbol{g} - \nu \boldsymbol{x}_0$ is a subgradient of F at $\boldsymbol{x}_0$: $\quad \forall \boldsymbol{x}_0 \in \mathbf{dom}\, f$,

$$f(\boldsymbol{x}) - \nu/2\|\boldsymbol{x}\|^2 \geq f(\boldsymbol{x}_0) - \nu/2\|\boldsymbol{x}_0\|^2 + \langle \boldsymbol{g} - \nu \boldsymbol{x}_0, \boldsymbol{x} - \boldsymbol{x}_0 \rangle.$$

- Re-arranging yields the proposition.

Minimizer of strong convex function

Theorem

Let f be a closed, $\nu-$strongly convex. Then f has a unique minimizer x^*, and

$$f(x) \geq f(x^*) + \frac{\nu}{2}\|x - x^*\|^2, \quad \forall x \in \mathbf{dom}\, f.$$

Proof:
See blackboard.

L-smoothness

Definition (L-smoothness).

A diffenrentiable function f is called *L-smooth* for some $L > 0$ if its gradient is L-Lipschitz:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbf{dom}\, f,$$

Remark: If f is twice diff., then f is L-smooth iff

$$\nabla^2 f(x) \preceq LI,\ \forall \boldsymbol{x} \in \mathbf{dom}\, f.$$

L-smoothness

Proposition

For a differentiable function f, consider the following statements:

(i) f is L-smooth (i.e., ∇f is L-Lipschitz)

(ii) $f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbf{dom}\, f.$

(iii) $\boldsymbol{x} \mapsto \frac{L}{2} \|\boldsymbol{x}\|^2 - f(\boldsymbol{x})$ is convex

It holds: (i) $\implies$ (ii) $\iff$ (iii).

If moreover f is convex, then (i) $\iff$ (ii) $\iff$ (iii).

We prove (i) $\implies$ (ii) on the next slide.

L-smoothness

- Let f be L-smooth, $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{dom}\, f$.
- From the fundamental theorem of calculus,

$$\begin{aligned} f(\boldsymbol{y}) &= f(\boldsymbol{x}) + \int_{t=0}^{1} \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y}-\boldsymbol{x})), \boldsymbol{y}-\boldsymbol{x} \rangle \, dt. \\ &= f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x} \rangle + \int_{t=0}^{1} \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y}-\boldsymbol{x})) - \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x} \rangle \, dt. \end{aligned}$$

L-smoothness

- Let f be L-smooth, $\boldsymbol{x}, \boldsymbol{y} \in \textbf{dom} f$.
- From the fundamental theorem of calculus,

$$\begin{aligned} f(\boldsymbol{y}) &= f(\boldsymbol{x}) + \int_{t=0}^{1} \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y}-\boldsymbol{x})), \boldsymbol{y}-\boldsymbol{x} \rangle \, dt. \\ &= f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x} \rangle + \int_{t=0}^{1} \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y}-\boldsymbol{x})) - \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x} \rangle \, dt. \end{aligned}$$

- Hence,

$$\begin{aligned} |f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x} \rangle| &= \left| \int_{t=0}^{1} \langle \nabla f(\boldsymbol{x} + t(\boldsymbol{y}-\boldsymbol{x})) - \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x} \rangle \, dt \right|. \\ \text{[Cauchy-Schwartz]} \quad &\leq \int_{t=0}^{1} \|\nabla f(\boldsymbol{x} + t(\boldsymbol{y}-\boldsymbol{x})) - \nabla f(\boldsymbol{x})\| \cdot \|\boldsymbol{y}-\boldsymbol{x}\| \, dt \\ \text{[L-smoothness]} \quad &\leq \int_{t=0}^{1} Lt\|\boldsymbol{y}-\boldsymbol{x}\| \cdot \|\boldsymbol{y}-\boldsymbol{x}\| \, dt = \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2. \end{aligned}$$

Outline

Prox operator

Definition (Prox operator).

Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a closed convex function. We define the proximal mapping of g by

$$\mathbf{prox}_g(x) := \underset{\boldsymbol{u}\in\mathbb{R}^n}{\operatorname{argmin}} \quad g(\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{u}\|^2.$$

The proximal operator generalizes the notion of *projection*:

- If C is convex set, define the convex indicator function

$$I_C(x) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{otherwise.} \end{cases}$$

- Then,

$$P_C(\boldsymbol{x}) = \underset{\boldsymbol{u}\in C}{\operatorname{argmin}} \|\boldsymbol{u}-\boldsymbol{x}\|^2 = \underset{\boldsymbol{u}}{\operatorname{argmin}} \|\boldsymbol{u}-\boldsymbol{x}\|^2 + I_C(\boldsymbol{u}) = \mathbf{prox}_{I_C}(\boldsymbol{x}).$$

Properties of prox

Theorem

Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a closed convex function. Then,

(i) $\mathbf{prox}_g(x) \in \mathbb{R}^n$ is well defined over $\mathbf{dom}\, g$.
(i.e., for all $x \in \mathbf{dom}\, g$, there is a unique minimizer).

(ii) $\boldsymbol{u} = \mathbf{prox}_g(\boldsymbol{x}) \iff \boldsymbol{x} - \boldsymbol{u} \in \partial g(\boldsymbol{u})$

(iii) $\boldsymbol{x}^*$ is a minimizer of $g \iff \boldsymbol{x}^* = \mathbf{prox}_g(\boldsymbol{x}^*)$.

Properties of prox

Theorem

Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a closed convex function. Then,

(i) $\mathbf{prox}_g(x) \in \mathbb{R}^n$ is well defined over $\mathbf{dom}\, g$.
(i.e., for all $x \in \mathbf{dom}\, g$, there is a unique minimizer).

(ii) $\boldsymbol{u} = \mathbf{prox}_g(x) \iff \boldsymbol{x} - \boldsymbol{u} \in \partial g(\boldsymbol{u})$

(iii) x^* is a minimizer of $g \iff x^* = \mathbf{prox}_g(x^*)$.

Proof:

(i) Let $h(\boldsymbol{u}) := g(\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{u}\|^2$. Then,

$$h(\boldsymbol{u}) - \frac{1}{2}\|\boldsymbol{u}\|^2 = g(\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{u}\|^2 - \frac{1}{2}\|\boldsymbol{u}\|^2 = g(\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{x}\|^2 - \boldsymbol{x}^T\boldsymbol{u}$$

is convex. So g is strongly convex with parameter $\nu = 1$ and has a single minimizer.

Properties of prox

Theorem

Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a closed convex function. Then,

(i) $\mathbf{prox}_g(x) \in \mathbb{R}^n$ is well defined over $\mathbf{dom}\, g$.
(i.e., for all $x \in \mathbf{dom}\, g$, there is a unique minimizer).

(ii) $\boldsymbol{u} = \mathbf{prox}_g(\boldsymbol{x}) \iff \boldsymbol{x} - \boldsymbol{u} \in \partial g(\boldsymbol{u})$

(iii) x^* is a minimizer of $g \iff x^* = \mathbf{prox}_g(x^*)$.

Proof:

(ii) Let $h(\boldsymbol{u}) := g(\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{x}\|^2$. The subdifferential or h is $\partial h(\boldsymbol{u}) = \partial g(\boldsymbol{u}) + \boldsymbol{u} - \boldsymbol{x}$. So $\boldsymbol{u}$ minimizes h iff

$$\mathbf{0} \in \partial g(\boldsymbol{u}) + \boldsymbol{u} - \boldsymbol{x} \iff \boldsymbol{x} - \boldsymbol{u} \in \partial g(\boldsymbol{u})$$

Properties of prox

Theorem

Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a closed convex function. Then,

(i) $\mathbf{prox}_g(x) \in \mathbb{R}^n$ is well defined over $\mathbf{dom}\, g$.
(i.e., for all $x \in \mathbf{dom}\, g$, there is a unique minimizer).

(ii) $\boldsymbol{u} = \mathbf{prox}_g(\boldsymbol{x}) \iff \boldsymbol{x} - \boldsymbol{u} \in \partial g(\boldsymbol{u})$

(iii) $\boldsymbol{x}^*$ is a minimizer of $g \iff \boldsymbol{x}^* = \mathbf{prox}_g(\boldsymbol{x}^*)$.

Proof:

(ii) Let $h(\boldsymbol{u}) := g(\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{x}\|^2$. The subdifferential or h is $\partial h(\boldsymbol{u}) = \partial g(\boldsymbol{u}) + \boldsymbol{u} - \boldsymbol{x}$. So $\boldsymbol{u}$ minimizes h iff

$$\mathbf{0} \in \partial g(\boldsymbol{u}) + \boldsymbol{u} - \boldsymbol{x} \iff \boldsymbol{x} - \boldsymbol{u} \in \partial g(\boldsymbol{u})$$

(iii) $\boldsymbol{x} = \mathbf{prox}_g(\boldsymbol{x}) \iff \boldsymbol{x} - \boldsymbol{x} = \mathbf{0} \in \partial g(\boldsymbol{x}) \iff \boldsymbol{x} \in \operatorname{argmin} g$.

Computing the prox

- In general, computing $\mathbf{prox}_g(x)$ can be as hard as minimizing g...
- Good news: for many functions, the prox operator can be computed efficiently, i.e. in $O(n)$ or $O(n\log(n))$, by using a closed-form formula, or by reducing to a one-dimensional problem.
- A catalog of known prox. operators can be found at `http://proximity-operator.net`
- Simple rule for separable sums:

$$\text{if } f(x_1,\ldots,x_n)=\sum_i f_i(x_i), \text{ then } \mathbf{prox}_f(x)=\begin{bmatrix}\mathbf{prox}_{f_1}(x_1)\\ \vdots \\ \mathbf{prox}_{f_n}(x_n)\end{bmatrix}$$

Example of proximal operators

We usually need to compute the proximal operator of a scaling $t \cdot g$ of a convex function g, for some $t > 0$.

$g(x)$	$\mathbf{prox}_{tg}(x)$
$\|\boldsymbol{x}\|_2$	$\left(1 - \dfrac{t}{\max(\|\boldsymbol{x}\|_2, t)}\right) \boldsymbol{x}$
$\boldsymbol{x}^T Q \boldsymbol{x} + \boldsymbol{p}^T \boldsymbol{x}$	$(tQ + I)^{-1}(\boldsymbol{x} - t\boldsymbol{p})$
$\|\boldsymbol{x}\|_1$	$[\lvert\boldsymbol{x}\rvert - t\mathbf{1}]_+ \odot \operatorname{sign}(\boldsymbol{x})$
$\sum_{i=1}^n [x_i]_+$	$\left[\left\lvert\boldsymbol{x} - \dfrac{t}{2}\mathbf{1}\right\rvert - \dfrac{t}{2}\mathbf{1}\right]_+ \odot \operatorname{sign}(\boldsymbol{x})$
$\sum_{i=1}^n x_i \log(x_i)$	$t\, W(t^{-1} e^{\frac{x}{t} - 1})$
$\max\limits_{i=1,\ldots,n} x_i$	$\{\min(x_i, s)\}_{i=1,\ldots,n}$ where s solves $\sum_i [\lvert x_i\rvert - s]_+ = t$

Example: Prox of ℓ_1-norm

- $f(x) = t\|x\|_1 = \sum_i t|x_i|$ is a separable sum, so we can focus on the 1-dimensional function $g : \mathbb{R} \to \mathbb{R}, x \to t|x|$ and apply the prox *elementwise*.

Example: Prox of ℓ_1-norm

- $f(x) = t\|x\|_1 = \sum_i t|x_i|$ is a separable sum, so we can focus on the 1-dimensional function $g : \mathbb{R} \to \mathbb{R}, x \to t|x|$ and apply the prox *elementwise*.
- $u^* = \mathbf{prox}_{x \mapsto t|x|}(x) \iff x - u^* \in \partial g(u^*) \iff$

$\big(x - u^* = -t \wedge u^* < 0\big)$ or $\big(x - u^* \in [-t, t] \wedge u^* = 0\big)$ or $\big(x - u^* = t \wedge u^* > 0\big)$.

Example: Prox of ℓ_1-norm

- $f(x) = t\|x\|_1 = \sum_i t|x_i|$ is a separable sum, so we can focus on the 1-dimensional function $g : \mathbb{R} \to \mathbb{R}, x \to t|x|$ and apply the prox *elementwise*.
- $u^* = \mathbf{prox}_{x \mapsto t|x|}(x) \iff x - u^* \in \partial g(u^*) \iff$

$\big(x - u^* = -t \wedge u^* < 0\big)$ or $\big(x - u^* \in [-t, t] \wedge u^* = 0\big)$ or $\big(x - u^* = t \wedge u^* > 0\big)$.

- We can solve this system by analyzing the sign of u^*:

$$\mathbf{prox}_{x \mapsto t|x|}(x) = u^* := \begin{cases} x + t & \text{if } x < -t \\ 0 & \text{if } x \in [-t, t] \\ x - t & \text{if } x > t \end{cases} = [|x| - t]_+ \operatorname{sign}(x),$$

Example: Prox of ℓ_1-norm

- $f(x) = t\|x\|_1 = \sum_i t|x_i|$ is a separable sum, so we can focus on the 1-dimensional function $g : \mathbb{R} \to \mathbb{R}, x \to t|x|$ and apply the prox *elementwise*.
- $u^* = \mathbf{prox}_{x \mapsto t|x|}(x) \iff x - u^* \in \partial g(u^*) \iff$

$$\big(x - u^* = -t \wedge u^* < 0\big) \text{ or } \big(x - u^* \in [-t, t] \wedge u^* = 0\big) \text{ or } \big(x - u^* = t \wedge u^* > 0\big).$$

- We can solve this system by analyzing the sign of u^*:

$$\mathbf{prox}_{x \mapsto t|x|}(x) = u^* := \begin{cases} x + t & \text{if } x < -t \\ 0 & \text{if } x \in [-t, t] \\ x - t & \text{if } x > t \end{cases} = [|x| - t]_+ \operatorname{sign}(x),$$

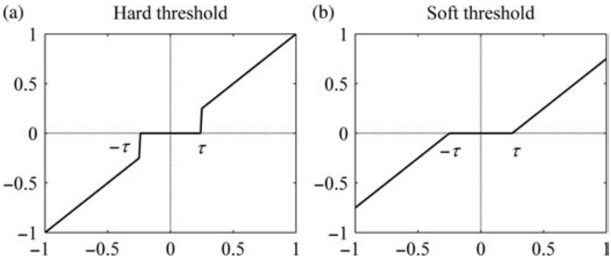
- Finally, apply the formula *componentwise*:

$$\mathbf{prox}_{tf}(\boldsymbol{x}) = \mathcal{T}_t(\boldsymbol{x}) := [|\boldsymbol{x}| - t\mathbf{1}]_+ \odot \operatorname{sign}(\boldsymbol{x}).$$

Prox of ℓ_1-norm: Soft thresholding

The proximal operator $\mathcal{T}_t(\boldsymbol{x})$ of $\boldsymbol{x} \mapsto t\|\boldsymbol{x}\|_1$ is called the *soft thresholding* operator (a level t).

$\mathcal{T}_\tau(\boldsymbol{x})$ acts on each coordinate as a thresholding operator that zeroes values $|x| < \tau$, but the function is shifted to make it continuous:



Outline

Composite model

Definition

From now on we consider a *Composite convex model*

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\textbf{minimize}}\ F(\boldsymbol{x}) := f(\boldsymbol{x}) + g(\boldsymbol{x}),$$

where

- $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex and L-smooth.
- $g : \mathbb{R}^n \to \mathbb{R}$ is closed, convex, possibly nonsmooth, but it has a *cheap proximal operator*.

Composite model

Definition

From now on we consider a *Composite convex model*

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\textbf{minimize}}\ F(\boldsymbol{x}) := f(\boldsymbol{x}) + g(\boldsymbol{x}),$$

where

- $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex and L-smooth.
- $g : \mathbb{R}^n \to \mathbb{R}$ is closed, convex, possibly nonsmooth, but it has a *cheap proximal operator*.

Special cases

- $g = 0$: Smooth convex, unconstrained optimization
- $g = I_C$: Minimization of f over the convex set C (for a "simple" convex set C such that the projection over C ($= \mathbf{prox}_{I_C}$) can be computed easily).

Basic idea

- $F(x) := f(x) + g(x)$, with f L-smooth, g proximable.

Basic idea

- $F(x) := f(x) + g(x)$, with f L-smooth, g proximable.
- f is $L-$smooth, so

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

Basic idea

- $F(x) := f(x) + g(x)$, with f L-smooth, g proximable.
- f is $L-$smooth, so

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

- At iteration k, we use this to obtain an overestimator of F around the current iterate $\boldsymbol{x}^{(k)}$: Given $0 < t_k \leq \frac{1}{L}$,

$$F(\boldsymbol{y}) \leq \hat{F}(\boldsymbol{y}) := f(\boldsymbol{x}^{(k)}) + \langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{y} - \boldsymbol{x}^{(k)} \rangle + \frac{1}{2t_k} \|\boldsymbol{y} - \boldsymbol{x}^{(k)}\|^2 + g(\boldsymbol{y})$$

Basic idea

- $F(x) := f(x) + g(x)$, with f L-smooth, g proximable.
- f is $L-$smooth, so

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

- At iteration k, we use this to obtain an overestimator of F around the current iterate $\boldsymbol{x}^{(k)}$: Given $0 < t_k \leq \frac{1}{L}$,

$$F(\boldsymbol{y}) \leq \hat{F}(\boldsymbol{y}) := f(\boldsymbol{x}^{(k)}) + \langle \nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{y} - \boldsymbol{x}^{(k)} \rangle + \frac{1}{2t_k} \|\boldsymbol{y} - \boldsymbol{x}^{(k)}\|^2 + g(\boldsymbol{y})$$

- The next iterate is determined by computing

$$\boldsymbol{x}^{(k+1)} := \operatorname*{argmin}_{\boldsymbol{y}} \ \hat{F}(\boldsymbol{y}).$$

This reduces to evaluating the $\mathbf{prox}_{t_k g}$ operator !

Proximal Gradient iteration

$$
\begin{aligned}
x^{(k+1)} :=& \operatorname*{argmin}_{y} \ \hat{F}(y) \\
=& \operatorname*{argmin}_{y} \ g(y) + y^T \nabla f(x^{(k)}) + \frac{1}{2t_k} \|y - x^{(k)}\|^2 \\
=& \operatorname*{argmin}_{y} \ g(y) + y^T \nabla f(x^{(k)}) + \frac{1}{2t_k} (\|y\|^2 - 2y^T x^{(k)}) \\
=& \operatorname*{argmin}_{y} \ t_k \, g(y) + \frac{1}{2} \|y\|^2 - y^T (x^{(k)} - t_k \nabla f(x^{(k)})) \\
=& \operatorname*{argmin}_{y} \ t_k \, g(y) + \frac{1}{2} \|y - (x^{(k)} - t_k \nabla f(x^{(k)}))\|^2 \\
=& \mathbf{prox}_{t_k g}(x^{(k)} - t_k \nabla f(x^{(k)}))
\end{aligned}
$$

Proximal Gradient iteration

$$
\begin{aligned}
x^{(k+1)} :=& \operatorname*{argmin}_{y} \ \hat{F}(y) \\
=& \operatorname*{argmin}_{y} \ g(y) + y^T \nabla f(x^{(k)}) + \frac{1}{2t_k} \|y - x^{(k)}\|^2 \\
=& \operatorname*{argmin}_{y} \ g(y) + y^T \nabla f(x^{(k)}) + \frac{1}{2t_k} (\|y\|^2 - 2y^T x^{(k)}) \\
=& \operatorname*{argmin}_{y} \ t_k \, g(y) + \frac{1}{2} \|y\|^2 - y^T (x^{(k)} - t_k \nabla f(x^{(k)})) \\
=& \operatorname*{argmin}_{y} \ t_k \, g(y) + \frac{1}{2} \|y - (x^{(k)} - t_k \nabla f(x^{(k)}))\|^2 \\
=& \mathbf{prox}_{t_k g}(x^{(k)} - t_k \nabla f(x^{(k)}))
\end{aligned}
$$

Definition Proximal Gradient iteration

For some step size $t > 0$, update $x^+ \leftarrow \mathbf{prox}_{tg}(x - t\nabla f(x))$.

Analysis of the proximal gradient method

Theorem (Prox-grad inequality).

Let $x \in \textbf{int dom} f$ denote the current iterate, and x^+ be the *next* iterate, obtained after a step of size $t > 0$, i.e.,

$$x^+ = \mathbf{prox}_{tg}(x - t\nabla f(x)).$$

If

$$f(x^+) \leq f(x) + \nabla f(x)^T (x^+ - x) + \frac{1}{2t}\|x^+ - x\|^2,$$

(so in particular, if $t \leq \frac{1}{L}$), then for all $\xi \in \mathbb{R}^n$ it holds:

$$F(\xi) - F(x^+) \geq \frac{1}{2t}\left(\|\xi - x^+\|^2 - \|\xi - x\|^2\right).$$

Proof of the prox-grad inequality (1/2)

- $x^+ = \mathbf{prox}_{tg}(x - t\nabla f(x)) \iff x - t\nabla f(x) - x^+ \in \partial(tg)(x^+)$
- So, by definition of a subgradient, for all $\xi \in \mathbb{R}^n$,

$$\begin{aligned} & tg(\xi) \geq tg(x^+) + \langle x - t\nabla f(x) - x^+, \xi - x^+\rangle \\ \iff & g(\xi) - g(x^+) \geq \frac{1}{t}\langle x - x^+, \xi - x^+\rangle - \nabla f(x)^T(\xi - x^+). \end{aligned}$$

Proof of the prox-grad inequality (1/2)

- $x^+ = \mathbf{prox}_{tg}(x - t\nabla f(x)) \iff x - t\nabla f(x) - x^+ \in \partial(tg)(x^+)$
- So, by definition of a subgradient, for all $\xi \in \mathbb{R}^n$,

$$\begin{aligned} & tg(\xi) \geq tg(x^+) + \langle x - t\nabla f(x) - x^+, \xi - x^+ \rangle \\ \iff & g(\xi) - g(x^+) \geq \frac{1}{t}\langle x - x^+, \xi - x^+ \rangle - \nabla f(x)^T(\xi - x^+). \end{aligned}$$

- And our assumption on the stepsize t can be rewritten as:

$$f(\xi) - f(x^+) \geq f(\xi) - f(x) - \nabla f(x)^T(x^+ - x) - \frac{1}{2t}\|x^+ - x\|^2$$

Proof of the prox-grad inequality (1/2)

- $x^+ = \mathbf{prox}_{tg}(x - t\nabla f(x)) \iff x - t\nabla f(x) - x^+ \in \partial(tg)(x^+)$
- So, by definition of a subgradient, for all $\xi \in \mathbb{R}^n$,

$$\begin{aligned} tg(\xi) &\geq tg(x^+) + \langle x - t\nabla f(x) - x^+, \xi - x^+ \rangle \\ \iff g(\xi) - g(x^+) &\geq \frac{1}{t}\langle x - x^+, \xi - x^+ \rangle - \nabla f(x)^T(\xi - x^+). \end{aligned}$$

- And our assumption on the stepsize t can be rewritten as:

$$f(\xi) - f(x^+) \geq f(\xi) - f(x) - \nabla f(x)^T(x^+ - x) - \frac{1}{2t}\|x^+ - x\|^2$$

- We sum the above two inequalities:

$$\begin{aligned} F(\xi) - F(x^+) \geq \underbrace{f(\xi) - f(x) - \nabla f(x)^T(\xi - x)}_{\epsilon_f(x,\xi)\geq 0} &+ \frac{1}{t}\langle x - x^+, \xi - x^+ \rangle \\ &- \frac{1}{2t}\|x^+ - x\|^2. \end{aligned}$$

Proof of the prox-grad inequality (2/2)

$$F(\xi) - F(x^+) \geq \underbrace{f(\xi) - f(x) - \nabla f(x)^T(\xi - x)}_{\epsilon_f(x,\xi) \geq 0} + \frac{1}{t}\langle x - x^+, \xi - x^+\rangle - \frac{1}{2t}\|x^+ - x\|^2.$$

So,

$$F(\xi) - F(x^+) \geq \frac{1}{2t}\left(2\langle x - x^+, \xi - x^+\rangle - \|x^+ - x\|^2\right).$$

Proof of the prox-grad inequality (2/2)

$$F(\xi) - F(x^+) \geq \underbrace{f(\xi) - f(x) - \nabla f(x)^T(\xi - x)}_{\epsilon_f(\boldsymbol{x},\xi)\geq 0} + \frac{1}{t}\langle \boldsymbol{x} - \boldsymbol{x}^+, \xi - \boldsymbol{x}^+\rangle - \frac{1}{2t}\|\boldsymbol{x}^+ - \boldsymbol{x}\|^2.$$

So,

$$F(\xi) - F(x^+) \geq \frac{1}{2t}\left(2\langle \boldsymbol{x} - \boldsymbol{x}^+, \xi - \boldsymbol{x}^+\rangle - \|\boldsymbol{x}^+ - \boldsymbol{x}\|^2\right).$$

Finally, we use the identity

$$\begin{aligned}\|\xi - \boldsymbol{x}\|^2 &= \|(\xi - \boldsymbol{x}^+) - (\boldsymbol{x} - \boldsymbol{x}^+)\|^2 \\ &= \|\xi - \boldsymbol{x}^+\|^2 + \|\boldsymbol{x} - \boldsymbol{x}^+\|^2 - 2\langle \xi - \boldsymbol{x}^+, \boldsymbol{c} - \boldsymbol{x}^+\rangle,\end{aligned}$$

Proof of the prox-grad inequality (2/2)

$$F(\xi) - F(x^+) \geq \underbrace{f(\xi) - f(x) - \nabla f(x)^T(\xi - x)}_{\epsilon_f(x,\xi)\geq 0} + \frac{1}{t}\langle x - x^+, \xi - x^+\rangle - \frac{1}{2t}\|x^+ - x\|^2.$$

So,

$$F(\xi) - F(x^+) \geq \frac{1}{2t}\left(2\langle x - x^+, \xi - x^+\rangle - \|x^+ - x\|^2\right).$$

Finally, we use the identity

$$\begin{aligned}\|\xi - x\|^2 &= \|(\xi - x^+) - (x - x^+)\|^2\\ &= \|\xi - x^+\|^2 + \|x - x^+\|^2 - 2\langle \xi - x^+, c - x^+\rangle,\end{aligned}$$

and we obtain the result:

$$F(\xi) - F(x^+) \geq \frac{1}{2t}\left(\|\xi - x^+\|^2 - \|\xi - x\|^2\right).$$

Sufficient decrease

Corollary

If the step size t is "well chosen" (i.e., it satifies the condition of the previous theorem), then

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

In particular, the proximal gradient method is a *descent method*.

Convergence Analysis

- We assume that $x^{(0)} \in \textbf{int dom}\, f$ and constant step sizes $t_k = \dfrac{1}{L}$ are used:

$$x^{(k+1)} := \mathbf{prox}_{\frac{1}{L}g}(x^{(k)} - \frac{1}{L}\nabla f(x^{(k)})).$$

- If L is unknown, one can use backtracking line search to find a step size t_k that satifies the condition of the previous theorem – then, similar analysis.

Theorem

For any optimal solution x^* of the composite convex optimization problem (**minimize** $f + g$),

$$F(x^{(k)}) - F(x^*) \leq \frac{L}{2k}\|x^{(0)} - x^*\|^2, \qquad \forall k \geq 1.$$

Convergence Analysis: Proof

- Prox-grad inequality at $\xi = \boldsymbol{x}^*$:

$$F(\boldsymbol{x}^*) - F(\boldsymbol{x}^{(i+1)}) \geq \frac{L}{2}\big(\|\boldsymbol{x}^* - \boldsymbol{x}^{(i+1)}\|^2 - \|\boldsymbol{x}^* - \boldsymbol{x}^{(i)}\|^2\big).$$

Convergence Analysis: Proof

- Prox-grad inequality at $\xi = \boldsymbol{x}^*$:

$$F(\boldsymbol{x}^*) - F(\boldsymbol{x}^{(i+1)}) \geq \frac{L}{2}\left(\|\boldsymbol{x}^* - \boldsymbol{x}^{(i+1)}\|^2 - \|\boldsymbol{x}^* - \boldsymbol{x}^{(i)}\|^2\right).$$

- Summing over $i = 0, \ldots, k-1$,

$$k\, F(\boldsymbol{x}^*) - \sum_{i=0}^{k-1} F(\boldsymbol{x}^{(i+1)}) \geq \frac{L}{2}\left(\|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\|^2 - \|\boldsymbol{x}^* - \boldsymbol{x}^{(0)}\|^2\right)$$

$$\Longrightarrow \sum_{i=1}^{k} F(\boldsymbol{x}^{(i)}) - k\, F(\boldsymbol{x}^*) \leq \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^{(0)}\|^2.$$

Convergence Analysis: Proof

- Prox-grad inequality at $\xi = \boldsymbol{x}^*$:

$$F(\boldsymbol{x}^*) - F(\boldsymbol{x}^{(i+1)}) \geq \frac{L}{2}\left(\|\boldsymbol{x}^* - \boldsymbol{x}^{(i+1)}\|^2 - \|\boldsymbol{x}^* - \boldsymbol{x}^{(i)}\|^2\right).$$

- Summing over $i = 0, \ldots, k-1$,

$$k\, F(\boldsymbol{x}^*) - \sum_{i=0}^{k-1} F(\boldsymbol{x}^{(i+1)}) \geq \frac{L}{2}\left(\|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\|^2 - \|\boldsymbol{x}^* - \boldsymbol{x}^{(0)}\|^2\right)$$

$$\implies \sum_{i=1}^{k} F(\boldsymbol{x}^{(i)}) - k\, F(\boldsymbol{x}^*) \leq \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^{(0)}\|^2.$$

- Since the algorithm is a descent method, $\displaystyle\sum_{i=1}^{k} F(\boldsymbol{x}^{(i)}) \geq kF(\boldsymbol{x}^{(k)})$.

Convergence Analysis: Proof

- Prox-grad inequality at $\xi = \boldsymbol{x}^*$:

$$F(\boldsymbol{x}^*) - F(\boldsymbol{x}^{(i+1)}) \geq \frac{L}{2}\big(\|\boldsymbol{x}^* - \boldsymbol{x}^{(i+1)}\|^2 - \|\boldsymbol{x}^* - \boldsymbol{x}^{(i)}\|^2\big).$$

- Summing over $i = 0, \ldots, k-1$,

$$k\,F(\boldsymbol{x}^*) - \sum_{i=0}^{k-1} F(\boldsymbol{x}^{(i+1)}) \geq \frac{L}{2}\big(\|\boldsymbol{x}^* - \boldsymbol{x}^{(k)}\|^2 - \|\boldsymbol{x}^* - \boldsymbol{x}^{(0)}\|^2\big)$$

$$\implies \sum_{i=1}^{k} F(\boldsymbol{x}^{(i)}) - k\,F(\boldsymbol{x}^*) \leq \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^{(0)}\|^2.$$

- Since the algorithm is a descent method, $\displaystyle\sum_{i=1}^{k} F(\boldsymbol{x}^{(i)}) \geq kF(\boldsymbol{x}^{(k)})$.

- Hence,

$$k(F(\boldsymbol{x}^{(k)}) - F(\boldsymbol{x}^*)) \leq \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^{(0)}\|^2.$$

Convergence Analysis

Theorem

For any optimal solution x^* of the composite convex optimization problem (**minimize** $f + g$),

$$F(\boldsymbol{x}^{(k)}) - F(\boldsymbol{x}^*) \leq \frac{L}{2k} \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|^2, \qquad \forall k \geq 1.$$

Remark

- It can also be shown that the sequence $(x^{(k)})_{k\in\mathbb{N}}$ converges to an optimal solution.
- This can NOT be considered as a polytime algorithm if ϵ is part of the input: $O(1/\epsilon)$ iterations required to find an ϵ-suboptimal solution, which is exponential w.r.t. input size $\langle\epsilon\rangle := |\log \epsilon|$.

Fast convergence for *strongly convex* functions

Now, consider a composite model (f, g) in which f is ν-strongly convex. Then, a *linear convergence rate* can be achieved (this time, we have a polytime algorihtm)

Theorem

If f is ν-strongly convex, then the proximal gradient method with constant step sizes ($t_k = \frac{1}{L}$) generates a sequence of points satisfying

(i) $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|^2 \leq \left(1 - \frac{\nu}{L}\right)^k \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|^2$;

(ii) $F(\boldsymbol{x}^{(k)}) - F(\boldsymbol{x}^*) \leq \frac{L}{2}\left(1 - \frac{\nu}{L}\right)^k \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|^2$,

where x^* denotes the *unique* optimal solution.

$\implies O\left(\frac{L}{\nu}\log(LR^2/\epsilon)\right)$ iterations to get ϵ-suboptimal solution.

Linear convergence: Proof sketch

- Recall proof of prox-grad inequality:

$$F(\xi) - F(\boldsymbol{x}^+) \geq \underbrace{\epsilon_f(\boldsymbol{x}, \xi)}_{\geq 0} + \frac{1}{2t}\left(\|\xi - \boldsymbol{x}^+\|^2 - \|\xi - \boldsymbol{x}\|^2\right).$$

Linear convergence: Proof sketch

- Recall proof of prox-grad inequality:

$$F(\xi) - F(x^+) \geq \underbrace{\epsilon_f(x, \xi)}_{\geq 0} + \frac{1}{2t}\left(\|\xi - x^+\|^2 - \|\xi - x\|^2\right).$$

- With f strongly convex, stronger bound: $\epsilon_f(x, \xi) \geq \dfrac{\nu}{2}\|\xi - x\|^2$.

Linear convergence: Proof sketch

- Recall proof of prox-grad inequality:

$$F(\xi) - F(\boldsymbol{x}^+) \geq \underbrace{\epsilon_f(\boldsymbol{x}, \xi)}_{\geq 0} + \frac{1}{2t}\left(\|\xi - \boldsymbol{x}^+\|^2 - \|\xi - \boldsymbol{x}\|^2\right).$$

- With f strongly convex, stronger bound: $\epsilon_f(\boldsymbol{x}, \xi) \geq \frac{\nu}{2}\|\xi - \boldsymbol{x}\|^2$.
- At $\xi = \xi^*$, with step size $t = 1/L$,

$$\begin{aligned} F(\boldsymbol{x}^*) - F(\boldsymbol{x}^+) &\geq \frac{L}{2}\left(\|\boldsymbol{x}^* - \boldsymbol{x}^+\|^2 - \|\boldsymbol{x}^* - \boldsymbol{x}\|^2\right) + \frac{\nu}{2}\|\boldsymbol{x}^* - \boldsymbol{x}\|^2 \\ &= \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^+\|^2 - \frac{L-\nu}{2}\|\boldsymbol{x}^* - \boldsymbol{x}\|^2. \end{aligned}$$

Linear convergence: Proof sketch

- Recall proof of prox-grad inequality:

$$F(\xi) - F(\boldsymbol{x}^+) \geq \underbrace{\epsilon_f(\boldsymbol{x},\xi)}_{\geq 0} + \frac{1}{2t}\left(\|\xi - \boldsymbol{x}^+\|^2 - \|\xi - \boldsymbol{x}\|^2\right).$$

- With f strongly convex, stronger bound: $\epsilon_f(\boldsymbol{x},\xi) \geq \frac{\nu}{2}\|\xi - \boldsymbol{x}\|^2$.
- At $\xi = \xi^*$, with step size $t = 1/L$,

$$\begin{aligned} F(\boldsymbol{x}^*) - F(\boldsymbol{x}^+) &\geq \frac{L}{2}\left(\|\boldsymbol{x}^* - \boldsymbol{x}^+\|^2 - \|\boldsymbol{x}^* - \boldsymbol{x}\|^2\right) + \frac{\nu}{2}\|\boldsymbol{x}^* - \boldsymbol{x}\|^2 \\ &= \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^+\|^2 - \frac{L-\nu}{2}\|\boldsymbol{x}^* - \boldsymbol{x}\|^2. \end{aligned}$$

- Then, we use $F(\boldsymbol{x}^*) - F(\boldsymbol{x}^+) \leq 0$:

$$\frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^+\|^2 \leq \frac{L-\nu}{2}\|\boldsymbol{x}^* - \boldsymbol{x}\|^2 \iff \|\boldsymbol{x}^* - \boldsymbol{x}^+\|^2 \leq \left(1 - \frac{\nu}{L}\right)\|\boldsymbol{x}^* - \boldsymbol{x}\|^2.$$

- The rest of the proof follows by easy induction.

Outline

History

- For smooth optimization, *accelerated* gradient methods were proposed by Nesterov [80's]
- Convergence in $\epsilon = O(1/k^2)$ instead of $O(1/k)$.
- The rate is *optimal* in some sense over the class of first-order methods
- Idea: Update $\boldsymbol{x}^{(k+1)}$ by taking a gradient step at point $\boldsymbol{y}^{(k)}$, where $\boldsymbol{y}^{(k)}$ is a well-chosen linear combination of previous 2 iterates, $\boldsymbol{x}^{(k)}$ and $\boldsymbol{x}^{(k-1)}$.
- Generalized to nonsmooth composite models by Beck & Tebeoulle [2009]. Method called FISTA for *fast iterative shrinkage-thresholding algorithm*, which describes the proximal gradient steps when $g(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$.

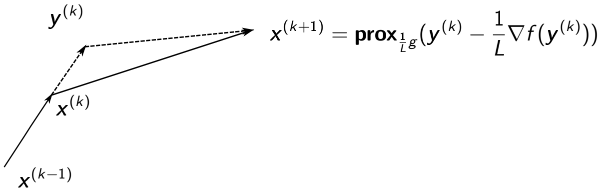
FISTA

FISTA (here, with constant step sizes $t_k = \frac{1}{L}, \forall k$)

Initialization: $y^{(0)} = x^{(0)} \in \textbf{int dom}\, f$, $\tau_0 = 1$.

For $k = 0, 1, 2, \ldots,$

1. $x^{(k+1)} = \textbf{prox}_{\frac{1}{L}g}\left(y^{(k)} - \frac{1}{L}\nabla f(y^{(k)})\right)$
2. $\tau_{k+1} = \dfrac{1 + \sqrt{1 + 4\tau_k^2}}{2}$
3. $y^{(k+1)} = x^{(k+1)} + \left(\dfrac{\tau_k - 1}{\tau_{k+1}}\right)(x^{(k+1)} - x^{(k)})$



FISTA

FISTA (here, with constant step sizes $t_k = \dfrac{1}{L}, \forall k$)

Initialization: $y^{(0)} = x^{(0)} \in \textbf{int dom}\, f$, $\tau_0 = 1$.

For $k = 0, 1, 2, \ldots,$

1. $x^{(k+1)} = \textbf{prox}_{\frac{1}{L} g}\left(y^{(k)} - \dfrac{1}{L} \nabla f(y^{(k)})\right)$
2. $\tau_{k+1} = \dfrac{1 + \sqrt{1 + 4\tau_k^2}}{2}$
3. $y^{(k+1)} = x^{(k+1)} + \left(\dfrac{\tau_k - 1}{\tau_{k+1}}\right)(x^{(k+1)} - x^{(k)})$

- In fact, τ_{k+1} solves the equation $\tau_{k+1}^2 - \tau_{k+1} = \tau_k^2$.
- Simple induction:

$$\tau_k \geq \frac{k+2}{2} \geq 1, \ \forall k \in \mathbb{N}.$$

Theorem

Consider the sequence of iterates $x^{(k)}$ generated by FISTA (with constant step sizes $t_k = \frac{1}{L}, \forall k$). Then, for any optimal solution x^* of the composite model (**minimize** $f + g$), it holds

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

Proof: See blackboard.

Example: Lasso regression (1/4)

In [1]:
```
#import packages
import numpy as np
import picos
%matplotlib inline
import matplotlib
import matplotlib.pyplot as plt
import time
```

The goal of this Notebook is to implement FISTA to solve the Lasso-regression problem

$$\min_x \|Ax - y\|^2 + \lambda \|x\|_1 .$$

In [2]:
```
#generate data
#In this example, y = Ax_0 + noise, where x_0 is sparse
m,n = 5000,1000
x0 = np.random.randn(n)
x0[:int(3*n/4)]=0
A = np.random.rand(m,n)
y = A.dot(x0) + 0.01 * np.random.rand(m)
lbda = 0.05
```

Example: Lasso regression (2/4)

Define Proximal (soft-thresholding) operator of $g = \|\cdot\|_1$

In [3]:
```
def prox_threshold(x,t):
    return np.maximum(0,np.abs(x)-t) * np.sign(x)
```

Compute Lipschitz constant

In [4]:
```
L = 2*lbda*max(np.linalg.svd(A)[1])**2
```

Proximal Gradient Method

In [ ]:
```
Niter = 10000
x = np.zeros(n)
prox_grad = []
for k in range(Niter):
    #compute value of objective function
    r = A.dot(x)-y
    prox_grad.append(lbda*np.linalg.norm(r)**2 + np.linalg.norm(x,1))
    #compute gradient of f
    grad = 2 * lbda * A.T.dot(r)
    #proximal step
    x = prox_threshold(x - 1./L * grad,1./L)
```

Example: Lasso regression (3/4)

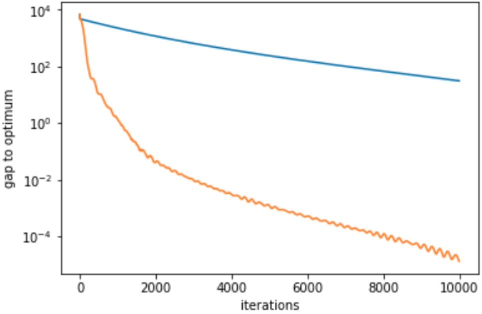
FISTA

In [ ]:
```
x_current = np.zeros(n)
y_current = np.zeros(n)
tau_current = 1.
fista = []
for k in range(Niter):
    #compute value of objective function
    fista.append(lbda*np.linalg.norm(A.dot(x_current)-y)**2 + np.linalg.norm(x_current,1)
    #compute gradient of f at y
    grad = 2 * lbda * A.T.dot(A.dot(y_current)-y)
    x_new = prox_threshold(y_current - 1./L * grad,1./L)
    #update tau and y
    tau_new = (1+(1+4*tau_current**2)**0.5)/2.
    y_current = x_new + (tau_current-1.)/tau_new * (x_new-x_current)
    #update current values
    x_current = x_new
    tau_current = tau_new
```

Example: Lasso regression (4/4)

```
In [18]: opt = min(min(prox_grad),min(fista))
         plt.semilogy(np.array(prox_grad)-opt)
         plt.semilogy(np.array(fista[:Niter])-opt)
         plt.xlabel("iterations")
         plt.ylabel("gap to optimum")
         print "time(proximal gradient)=",t_proxgrad
         print "time(FISTA)=",t_fista

time(proximal gradient)= 44.049719
time(FISTA)= 59.7696934545
```



Example: Lasso regression

- On this example, FISTA $\gg$ standard proximal gradient
- But FISTA *is not* a descent method
- The upper bounds for the gap $\delta_k \leq \dfrac{LR^2}{2k}$ and $\delta_k \leq \dfrac{2LR^2}{(k+1)^2}$ are very pessimistic:
  After $k = 10^4$ iterations, assuming the exact value of $R = \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|$ is known,

Algorithm	δ_k	upper bound
Proximal gradient	30.77	1421.41
FISTA	$1.29 \cdot 10^{-5}$	0.5684

- In practice, we can use much better duality bound on δ_k as stopping criterion.

Outline

1. Gradient & Subgradient methods
2. Strong convexity and L-smoothness
3. The proximal operator
4. The proximal gradient method
5. The FISTA accelerated method
6. **Optimality of accelerated gradient methods**

$\Omega(1/k^2)$ lower bound

The following result basically states that $O(1/k^2)$ is the best convergence rate we can hope for in the class of first-order methods:

Theorem

There exists a function $f:\mathbb{R}^{2k+1}\to\mathbb{R}$ which is twice differentiable and L-smooth, such that for any sequence $(x^{(i)})_{i\in\mathbb{N}}$ satisfying

$$x^{(i+1)} \in x^{(0)} + \operatorname{span}(\nabla f(x^{(0)}),\ldots,\nabla f(x^{(i)})),\ \forall i \in \mathbb{N},$$

it holds

$$f(x^{(k)}) - f(x^*) \geq \frac{3L\|x^{(0)}-x^*\|^2}{32(k+1)^2}.$$

$\Omega(1/k^2)$ lower bound: Proof sketch

$$f_k(\boldsymbol{x}) := \frac{L}{4}\left(\frac{1}{2}\boldsymbol{x}^T A \boldsymbol{x} - \boldsymbol{e}_1^T \boldsymbol{x}\right), \text{ where } A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{S}^k.$$

Assume (w.l.o.g.) that $\boldsymbol{x}^{(0)} = \mathbf{0}$. We can show that

$\Omega(1/k^2)$ lower bound: Proof sketch

$$f_k(\boldsymbol{x}) := \frac{L}{4}(\frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} - \boldsymbol{e}_1^T\boldsymbol{x}), \text{ where } A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{S}^k.$$

Assume (w.l.o.g.) that $\boldsymbol{x}^{(0)} = \mathbf{0}$. We can show that

- $A \succeq 0$ and $\lambda_{\max}(A) \leq 4$, hence f_k is convex and L-smooth, $\forall k$.

$\Omega(1/k^2)$ lower bound: Proof sketch

$$f_k(\boldsymbol{x}) := \frac{L}{4}\left(\frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} - \boldsymbol{e}_1^T\boldsymbol{x}\right), \text{ where } A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{S}^k.$$

Assume (w.l.o.g.) that $\boldsymbol{x}^{(0)} = \mathbf{0}$. We can show that

- $A \succeq 0$ and $\lambda_{\max}(A) \leq 4$, hence f_k is convex and L-smooth, $\forall k$.
- f_k is minimized over $\mathbb{R}^k$ at $\boldsymbol{x}^* = A^{-1}\boldsymbol{e}_1$, and

$$f_k(\boldsymbol{x}^*) = -\frac{L}{8}\boldsymbol{e}_1^T A^{-1}\boldsymbol{e}_1 = -\frac{L}{8}\left(1 - \frac{1}{k+1}\right).$$

$\Omega(1/k^2)$ lower bound: Proof sketch

$$f_k(\boldsymbol{x}) := \frac{L}{4}(\frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} - \boldsymbol{e}_1^T\boldsymbol{x}), \text{ where } A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{S}^k.$$

Assume (w.l.o.g.) that $\boldsymbol{x}^{(0)} = \mathbf{0}$. We can show that

- $A \succeq 0$ and $\lambda_{\max}(A) \leq 4$, hence f_k is convex and L-smooth, $\forall k$.
- f_k is minimized over $\mathbb{R}^k$ at $\boldsymbol{x}^* = A^{-1}\boldsymbol{e}_1$, and

$$f_k(\boldsymbol{x}^*) = -\frac{L}{8}\boldsymbol{e}_1^T A^{-1}\boldsymbol{e}_1 = -\frac{L}{8}(1 - \frac{1}{k+1}).$$

- Let $f = f_{2k+1}$. A simple induction shows that for all $i < 2k+1$,

$$\text{span}(\nabla f(\boldsymbol{x}^{(0)}), \dots, \nabla f(\boldsymbol{x}^{(i)})) \subseteq \text{span}(\boldsymbol{e}_1, \dots, \boldsymbol{e}_{i+1}).$$

$\Omega(1/k^2)$ lower bound: Proof sketch

$$f_k(\boldsymbol{x}) := \frac{L}{4}(\frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} - \boldsymbol{e}_1^T\boldsymbol{x}), \text{ where } A = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & & & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{S}^k.$$

Assume (w.l.o.g.) that $\boldsymbol{x}^{(0)} = \boldsymbol{0}$. We can show that

- $A \succeq 0$ and $\lambda_{\max}(A) \leq 4$, hence f_k is convex and L-smooth, $\forall k$.
- f_k is minimized over $\mathbb{R}^k$ at $\boldsymbol{x}^* = A^{-1}\boldsymbol{e}_1$, and

$$f_k(\boldsymbol{x}^*) = -\frac{L}{8}\boldsymbol{e}_1^T A^{-1}\boldsymbol{e}_1 = -\frac{L}{8}(1 - \frac{1}{k+1}).$$

- Let $f = f_{2k+1}$. A simple induction shows that for all $i < 2k+1$,

$$\operatorname{span}(\nabla f(\boldsymbol{x}^{(0)}), \ldots, \nabla f(\boldsymbol{x}^{(i)})) \subseteq \operatorname{span}(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{i+1}).$$

- So, $f(\boldsymbol{x}^{(k)}) \geq \inf_{\boldsymbol{z}} f_k(\boldsymbol{z}) = -\frac{L}{8}(1 - \frac{1}{k+1})$ and $f(\boldsymbol{x}^*) = -\frac{L}{8}(1 - \frac{1}{2k+2})$

$\Omega(1/k^2)$ lower bound: Proof sketch

$$f_k(\boldsymbol{x}) := \frac{L}{4}\left(\frac{1}{2}\boldsymbol{x}^T A \boldsymbol{x} - \boldsymbol{e}_1^T \boldsymbol{x}\right), \text{ where } A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{S}^k.$$

Assume (w.l.o.g.) that $\boldsymbol{x}^{(0)} = \boldsymbol{0}$. We can show that

- $A \succeq 0$ and $\lambda_{\max}(A) \leq 4$, hence f_k is convex and L-smooth, $\forall k$.
- f_k is minimized over $\mathbb{R}^k$ at $\boldsymbol{x}^* = A^{-1}\boldsymbol{e}_1$, and

$$f_k(\boldsymbol{x}^*) = -\frac{L}{8}\boldsymbol{e}_1^T A^{-1} \boldsymbol{e}_1 = -\frac{L}{8}\left(1 - \frac{1}{k+1}\right).$$

- Let $f = f_{2k+1}$. A simple induction shows that for all $i < 2k+1$,

$$\operatorname{span}(\nabla f(\boldsymbol{x}^{(0)}), \ldots, \nabla f(\boldsymbol{x}^{(i)})) \subseteq \operatorname{span}(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{i+1}).$$

- So, $f(\boldsymbol{x}^{(k)}) \geq \inf_{\boldsymbol{z}} f_k(\boldsymbol{z}) = -\frac{L}{8}\left(1 - \frac{1}{k+1}\right)$ and $f(\boldsymbol{x}^*) = -\frac{L}{8}\left(1 - \frac{1}{2k+2}\right)$
- Finally, we can bound $\|\boldsymbol{x}^*\|^2 \leq \frac{2}{3}(k+1)$. Putting all together yields the desired bound.