Convex Optimization and Applications
12 - First Order Methods

Guillaume Sagnol
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.
- → preference given to first-order algorithms, that quickly converge to a reasonably good solution.
Examples (1/2)

- **Lasso regression**

\[
\minimize_{\theta \in \mathbb{R}^n} \|X\theta - y\|^2 + \lambda \|\theta\|_1,
\]

- **Soft-margin SVM**

\[
\minimize_{w \in \mathbb{R}^n, b \in \mathbb{R}} \sum_{i=1}^{m} \max(0, 1 - y_i(w^T x_i - b)) + \lambda \|w\|^2.
\]

- **D-optimal design**

\[
\maximize_{w \in \mathbb{R}^n} \det^{\frac{1}{n}} \left( \sum_{i=1}^{m} w_i x_i x_i^T \right)
\]

\[
\text{s.t. } w \geq 0, \sum_i w_i = 1.
\]
Examples (2/2)

■ Low-rank Matrix completion

\[
\min_{Y \in \mathbb{R}^{m \times n}} \sum_{(i,j) \in \Omega} (X_{ij} - Y_{ij})^2 + \lambda \|Y\|_*,
\]

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

■ Total-Variation denoising

\[
\min_{Y \in \mathbb{R}^{m \times n}} \|X - Y\|_F^2 + \lambda \text{TV}(Y),
\]

The total-variation \( \text{TV}(Y) := \sum_{i,j} \left\| \begin{bmatrix} y_{i+1,j} - y_{i,j} \\ y_{i,j+1} - y_{i,j} \end{bmatrix} \right\|_2 \)

penalizes the pixels with a high local variation (noise).
Outline

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
Gradient descent

- A first order method is an algorithm to minimize a function $F$, that only uses first-order 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 $g$ is a subgradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \text{dom } f$, if

\[ f(z) \geq f(x) + g^T(z - x), \quad \forall z \in \text{dom } f. \]

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

The subdifferential of $f$ at $x$ is the set of all subgradients:

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

Proposition

(i) \( \partial f(x) \) is always a closed convex set.
(ii) \( f \) convex \( \implies \partial f(x) \neq \emptyset, \forall x \in \text{int dom } f \).
(iii) Let \( f \) be convex and \( x \in \text{int dom } f \). Then,

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

For (i), we use that

\[
\partial f(x) = \bigcap_{z \in \text{dom } f} \{ g : f(z) \geq f(x) + g^T(z - x) \}
\]

is an intersection of halfspaces.

(ii) follows from the supporting hyperplane theorem.
Example

Let $f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto |x|$. 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 $0 \in \partial f(x^*)$.

**Proof:**

$$0 \in \partial f(x^*) \iff f(z) \geq f(x^*) + 0^T(z - x^*), \forall z \in \text{dom } f.$$
Calculus rules for subdifferentials

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

- Nonnegative scaling:

\[ \partial(\alpha f)(x) = \alpha \partial f(x), \text{ for all } \alpha \geq 0. \]
Calculus rules for subdifferentials

Let \( f, f_1, \ldots, f_m \) be \textbf{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).
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:**
  \[
  \partial \left( z \mapsto f(Az + b) \right)(x) = A^T \partial f(Ax + b).
  \]
Calculus rules for subdifferentials

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

■ Nonnegative scaling:

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

■ 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:

$$
\partial(z \mapsto f(Az + b))(x) = A^T \partial f(Ax + b).
$$

■ Pointwise maximum:

Let $g(x) = \max_{i=1,\ldots,m} f_i(x)$. Then, 

$$
\partial g(x) = \text{conv} \left( \bigcup_{j \in A(x)} \partial f_j(x) \right),
$$

where $A(x)$ is the set of active functions at $x$, i.e.,

$$
A(x) := \{ j \in [m] : f_j(x) = g(x) \}.
$$

(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:

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

A few properties of this algorithm:

- Not a *descent method* (we can have $F(x^{(k)}) > F(x^{(k-1)})$, 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:

\[
x^{(k)} = x^{(k-1)} - t_k g^{(k-1)}, \quad \text{for some } g^{(k-1)} \in \partial F(x^{(k-1)}).
\]

A few properties of this algorithm:
- Not a descent method (we can have \( F(x^{(k)}) > F(x^{(k-1)}) \), 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:

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

A few properties of this algorithm:

- Not a descent method (we can have $F(x^{(k)}) > F(x^{(k-1)})$, 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:

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

A few properties of this algorithm:

- Not a descent method (we can have $F(x^{(k)}) > F(x^{(k-1)})$, 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 $f(x^{(k)}_{\text{best}}) \leq f(x^*) + O\left(\frac{1}{\sqrt{k}}\right)$. 
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
**Strong convexity**

**Definition (\(\nu\)-strong convexity).**

\(f\) is \(\nu\)-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 \(\nu\)-strongly convex iff

\[
\nabla^2 f(x) \succeq \nu I, \; \forall x \in \text{dom } f.
\]

G. Sagnol 12 - First Order Methods for Large Scale Problems 14 / 51
Strong convexity

**Definition (\(\nu\)-strong convexity).**

\(f\) is \(\nu\)-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 \(\nu\)-strongly convex iff

\[\nabla^2 f(x) \succeq \nu I, \quad \forall x \in \text{dom } f.\]

**Proposition**

Let \(f\) be \(\nu\)-strongly convex. Then, \(\forall x_0 \in \text{dom } \partial f, \forall g \in \partial f(x_0),\)

\[f(x) \geq f(x_0) + \langle g, x - x_0 \rangle + \frac{\nu}{2}\|x - x_0\|^2, \quad \forall x \in \text{dom } f.\]

**Remark:** The converse statement is also true.
Strong convexity

Proposition

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

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

Proof:

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

  $$g \in \partial f(x_0) = \partial F(x_0) + \frac{\nu}{2} \nabla (x \mapsto \|x\|^2) = \partial F(x_0) + \nu x_0$$

- So $g - \nu x_0$ is a subgradient of $F$ at $x_0$: $\forall x_0 \in \text{dom } f$,

  $$f(x) - \nu/2 \|x\|^2 \geq f(x_0) - \nu/2 \|x_0\|^2 + \langle g - \nu x_0, x - x_0 \rangle.$$

- Re-arranging yields the proposition.
Theorem

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

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

Proof:
See blackboard.
### Definition (L-smoothness).

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

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

### Remark:

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

$$
\nabla^2 f(x) \preceq L I, \quad \forall x \in \text{dom } f.
$$
**Proposition**

For a differentiable function $f$, consider the following statements:

(i) $f$ is $L$-smooth (i.e., $\nabla f$ is $L$-Lipschitz)

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

(iii) $x \mapsto \frac{L}{2} \|x\|^2 - f(x)$ 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.
\textbf{L-smoothness}

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

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

$$= f(x) + \langle \nabla f(x), y - x \rangle + \int_{t=0}^{1} \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle \, dt.$$
\(L\)-smoothness

- Let \(f\) be \(L\)-smooth, \(x, y \in \text{dom } f\).
- From the fundamental theorem of calculus,

\[
f(y) = f(x) + \int_{t=0}^{1} \langle \nabla f(x + t(y - x)), y - x \rangle \ dt.
\]

\[
= f(x) + \langle \nabla f(x), y - x \rangle + \int_{t=0}^{1} \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle \ dt.
\]

- Hence,

\[
|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| = \left| \int_{t=0}^{1} \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle \ dt \right|.
\]

[Cauchy-Schwartz]

\[
\leq \int_{t=0}^{1} \| \nabla f(x + t(y - x)) - \nabla f(x) \| \cdot \| y - x \| \ dt
\]

[L-smoothness]

\[
\leq \int_{t=0}^{1} Lt \| y - x \| \cdot \| y - x \| \ dt = \frac{L}{2} \| x - y \|^2.
\]
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
Prox operator

Definition (Prox operator).

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

$$\text{prox}_g(x) := \arg\min_{u \in \mathbb{R}^n} g(u) + \frac{1}{2} \|x - u\|^2.$$ 

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,

$$P_C(x) = \arg\min_{u \in C} \|u - x\|^2 = \arg\min_u \|u - x\|^2 + I_C(u) = \text{prox}_{I_C}(x).$$
Properties of prox

**Theorem**

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

(i) \( \text{prox}_g(x) \in \mathbb{R}^n \) is well defined over \( \text{dom} \ g \).

(i.e., for all \( x \in \text{dom} \ g \), there is a unique minimizer).

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

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

**Theorem**

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

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

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

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

**Proof:**

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

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

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

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

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

**Proof:**

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

\[
0 \in \partial g(u) + u - x \iff x - u \in \partial g(u)
\]
Properties of prox

Theorem

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

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

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

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

Proof:

(ii) Let $h(u) := g(u) + \frac{1}{2} \|u - x\|^2$. The subdifferential of $h$ is
    $\partial h(u) = \partial g(u) + u - x$. So $u$ minimizes $h$ iff
    $$0 \in \partial g(u) + u - x \iff x - u \in \partial g(u)$$

(iii) $x = \text{prox}_g(x) \iff x - x = 0 \in \partial g(x) \iff x \in \text{argmin} g.$
Computing the prox

- In general, computing $\text{prox}_g(x)$ 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:

  \[
  \text{if } f(x_1, \ldots, x_n) = \sum_i f_i(x_i), \text{ then } \text{prox}_f(x) = \begin{bmatrix}
  \text{prox}_{f_1}(x_1) \\
  \vdots \\
  \text{prox}_{f_n}(x_n)
  \end{bmatrix}
  \]
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 $t > 0$.

<table>
<thead>
<tr>
<th>$g(x)$</th>
<th>$\text{prox}_{tg}(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$|x|_2$</td>
<td>$\left(1 - \frac{t}{\max(|x|_2, t)}\right)x$</td>
</tr>
<tr>
<td>$x^TQx + p^Tx$</td>
<td>$(tQ + I)^{-1}(x - tp)$</td>
</tr>
<tr>
<td>$|x|_1$</td>
<td>$[</td>
</tr>
<tr>
<td>$\sum_{i=1}^n [x_i]_+$</td>
<td>$\left[</td>
</tr>
<tr>
<td>$\sum_{i=1}^n x_i \log(x_i)$</td>
<td>$t W(t^{-1}e^{\frac{x}{t} - 1})$</td>
</tr>
<tr>
<td>$\max_{i=1,\ldots,n} x_i$</td>
<td>${\min(x_i, s)}_{i=1,\ldots,n}$ where $s$ solves $\sum_i[</td>
</tr>
</tbody>
</table>

where $s$ solves $\sum_i[|x_i| - s]_+ = t$
Example: Prox of $\ell_1$-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 \textit{elementwise}. 

\[
\text{prox}_{t|x|}(x) := \begin{cases} 
  x + t & \text{if } x < -t \\
  0 & \text{if } x \in [-t, t] \\
  x - t & \text{if } x > t
\end{cases}
\]

Finally, apply the formula componentwise:

\[
\text{prox}_{tf}(x) = T_t(x) := \left[ |x| - t \right]_+ \circ \text{sign}(x)
\]
Example: Prox of $\ell_1$-norm

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

- $u^* = \text{prox}_{x \mapsto t|x|}(x) \iff x - u^* \in \partial g(u^*) \iff (x - u^* = -t \land u^* < 0)$ or $(x - u^* \in [-t, t] \land u^* = 0)$ or $(x - u^* = t \land u^* > 0)$. 

- Finally, apply the formula componentwise: $\text{prox}_t f(x) = T_t(x) := |x| - t 1_{\text{sign}(x)}$. 

G. Sagnol 12 - First Order Methods for Large Scale Problems 25 / 51
Example: Prox of $\ell_1$-norm

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

- $u^* = \text{prox}_{x \mapsto t|x|}(x) \iff x - u^* \in \partial g(u^*) \iff (x-u^* = -t \land u^* < 0)$ or $(x-u^* \in [-t, t] \land u^* = 0)$ or $(x-u^* = t \land u^* > 0)$.

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

$$\text{prox}_{x \mapsto t|x|}(x) = u^* := \begin{cases} 
  x + t & \text{if } x < -t \\
  0 & \text{if } x \in [-t, t] = \lfloor |x| - t \rfloor_+ \text{sign}(x), \\
  x - t & \text{if } x > t
\end{cases}$$
Example: Prox of \( \ell_1 \)-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.

- \( u^* = \text{prox}_{x \mapsto t|x|}(x) \iff x - u^* \in \partial g(u^*) \iff (x - u^* = -t \land u^* < 0) \) or \( (x - u^* \in [-t, t] \land u^* = 0) \) or \( (x - u^* = t \land u^* > 0) \).

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

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

- Finally, apply the formula componentwise:

  \[
  \text{prox}_{tf}(x) = T_t(x) := [\|x\| - t1]_+ \odot \text{sign}(x).
  \]
Prox of $\ell_1$-norm: Soft thresholding

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

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

**Definition**

From now on we consider a *Composite convex model*

\[
\minimize_{x \in \mathbb{R}^n} F(x) := f(x) + g(x),
\]

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 \( C \) (i.e., \( \text{prox} I_C \)) can be computed easily).
Composite model

Definition

From now on we consider a Composite convex model

\[
\begin{align*}
\text{minimize } & F(x) := f(x) + g(x), \\
\text{where } & f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\} \text{ is convex and } L\text{-smooth.} \\
& g : \mathbb{R}^n \to \mathbb{R} \text{ is closed, convex, possibly nonsmooth, but it has a cheap proximal operator.}
\end{align*}
\]

Special cases

- \( g = 0 \): Smooth convex, unconstrained optimization
- \( g = l_C \): Minimization of \( f \) over the convex set \( C \) (for a “simple” convex set \( C \) such that the projection over \( C \) (\( = \text{prox}_{l_C} \)) 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

\[
  f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \| y - x \|^2, \quad \forall x, y \in \text{dom} \ f.
\]
Basic idea

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

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

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

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

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

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

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

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

- The next iterate is determined by computing

$$x^{(k+1)} := \arg\min_y \hat{F}(y).$$

This reduces to evaluating the $\text{prox}_{t_k g}$ operator!
Proximal Gradient iteration

\[ x^{(k+1)} := \arg\min_y \hat{F}(y) \]

\[ = \arg\min_y g(y) + y^T \nabla f(x^{(k)}) + \frac{1}{2t_k} \|y - x^{(k)}\|^2 \]

\[ = \arg\min_y g(y) + y^T \nabla f(x^{(k)}) + \frac{1}{2t_k} (\|y\|^2 - 2y^T x^{(k)}) \]

\[ = \arg\min_y t_k g(y) + \frac{1}{2} \|y\|^2 - y^T (x^{(k)} - t_k \nabla f(x^{(k)})) \]

\[ = \arg\min_y t_k g(y) + \frac{1}{2} \|y - (x^{(k)} - t_k \nabla f(x^{(k)}))\|^2 \]

\[ = \text{prox}_{t_k g}(x^{(k)} - t_k \nabla f(x^{(k)})) \]
Proximal Gradient iteration

\[ x^{(k+1)} := \arg\min_y \hat{F}(y) \]

\[ = \arg\min_y g(y) + y^T \nabla f(x^{(k)}) + \frac{1}{2t_k} \|y - x^{(k)}\|^2 \]

\[ = \arg\min_y g(y) + y^T \nabla f(x^{(k)}) + \frac{1}{2t_k} (\|y\|^2 - 2y^T x^{(k)}) \]

\[ = \arg\min_y t_k g(y) + \frac{1}{2} \|y\|^2 - y^T (x^{(k)} - t_k \nabla f(x^{(k)})) \]

\[ = \arg\min_y t_k g(y) + \frac{1}{2} \|y - (x^{(k)} - t_k \nabla f(x^{(k)}))\|^2 \]

\[ = \text{prox}_{t_k g}(x^{(k)} - t_k \nabla f(x^{(k)})) \]

**Definition Proximal Gradient iteration**

For some step size \( t > 0 \), update \( x^+ \leftarrow \text{prox}_{t g}(x - t \nabla f(x)) \).
Analysis of the proximal gradient method

Theorem (Prox-grad inequality).

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

\[
x^+ = \text{prox}_{tg}(x - t\nabla f(x)).
\]

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:

\[
F(\xi) - F(x^+) \geq \frac{1}{2t}(\|\xi - x^+\|^2 - \|\xi - x\|^2).
\]
Proof of the prox-grad inequality (1/2)

\[ x^+ = \text{prox}_{t g}(x - t \nabla f(x)) \iff x - t \nabla f(x) - x^+ \in \partial(t g)(x^+) \]

So, by definition of a subgradient, for all \( \xi \in \mathbb{R}^n \),

\[ tg(\xi) \geq tg(x^+) + \langle x - t \nabla f(x) - x^+, \xi - x^+ \rangle \]

\[ \iff g(\xi) - g(x^+) \geq \frac{1}{t} \langle x - x^+, \xi - x^+ \rangle - \nabla f(x)^T (\xi - x^+) \]
Proof of the prox-grad inequality (1/2)

- \( x^+ = \text{prox}_{tg}(x - t\nabla f(x)) \iff x - t\nabla f(x) - x^+ \in \partial(tg)(x^+) \)

- So, by definition of a subgradient, for all \( \xi \in \mathbb{R}^n \),

\[
tg(\xi) \geq tg(x^+) + \langle x - t\nabla f(x) - x^+, \xi - x^+ \rangle
\]

\[
\iff g(\xi) - g(x^+) \geq \frac{1}{t} \langle x - x^+, \xi - x^+ \rangle - \nabla f(x)^T(\xi - x^+).
\]

- 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)

- \( x^+ = \text{prox}_{tg}(x - t\nabla f(x)) \iff x - t\nabla f(x) - x^+ \in \partial(tg)(x^+) \)

- So, by definition of a subgradient, for all \( \xi \in \mathbb{R}^n \),
  \[
  tg(\xi) \geq tg(x^+) + \langle x - t\nabla f(x) - x^+, \xi - x^+ \rangle \\
  \iff g(\xi) - g(x^+) \geq \frac{1}{t} \langle x - x^+, \xi - x^+ \rangle - \nabla f(x)^T(\xi - x^+). 
  \]

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

\[ F(\xi) - F(x^+) \geq f(\xi) - f(x) - \nabla f(x)^T(\xi - x) + \frac{1}{t} \langle x - x^+, \xi - x^+ \rangle - \frac{1}{2t} \|x^+ - x\|^2. \]

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

\[ \|\xi - x\|^2 = \| (\xi - x^+) - (x - x^+) \|^2 \]
\[ = \|\xi - x^+\|^2 + \|x - x^+\|^2 - 2 \langle \xi - x^+, c - x^+ \rangle, \]
Proof of the prox-grad inequality (2/2)

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

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

\[
\|\xi - x\|^2 = \|(\xi - x^+) - (x - x^+)\|^2
= \|\xi - x^+\|^2 + \|x - x^+\|^2 - 2\langle \xi - x^+, c - x^+ \rangle,
\]

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 satisfies 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 \textit{descent method}. 
Convergence Analysis

- We assume that $x^{(0)} \in \text{int dom } f$ and constant step sizes $t_k = \frac{1}{L}$ are used:
  
  $$x^{(k+1)} := \text{prox}_{\frac{1}{L}g}(x^{(k)} - \frac{1}{L}\nabla f(x^{(k)})).$$

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

Theorem

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

$$F(x^{(k)}) - F(x^*) \leq \frac{L}{2k}\|x^{(0)} - x^*\|^2, \quad \forall k \geq 1.$$
Convergence Analysis: Proof

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

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

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

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

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

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

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

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

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

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

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

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

- Since the algorithm is a descent method, $\sum_{i=1}^{k} F(x^{(i)}) \geq k F(x^{(k)})$. 
Convergence Analysis: Proof

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

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

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

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

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

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

- Hence,

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

Theorem

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

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

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 $\epsilon$ is part of the input: $O(1/\epsilon)$ iterations required to find an $\epsilon$-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 \(\nu\)-strongly convex. Then, a *linear convergence rate* can be achieved (this time, we have a polytime algorithm)

**Theorem**

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

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

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

where \(x^*\) denotes the *unique* optimal solution.

\[\Rightarrow \quad O\left(\frac{L}{\nu} \log(LR^2/\epsilon)\right) \text{ iterations to get } \epsilon\text{-suboptimal solution.}\]
Linear convergence: Proof sketch

- Recall proof of prox-grad inequality:

\[ F(\xi) - F(x^+) \geq \epsilon_f(x, \xi) + \frac{1}{2t} (\|\xi - x^+\|^2 - \|\xi - x\|^2). \]
Linear convergence: Proof sketch

- Recall proof of prox-grad inequality:

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

- With \( f \) strongly convex, stronger bound: \( \epsilon_f(x, \xi) \geq \frac{\nu}{2} \|\xi - x\|^2 \).
Linear convergence: Proof sketch

- Recall proof of prox-grad inequality:

\[ F(\xi) - F(x^+) \geq \epsilon_f(x, \xi) + \frac{1}{2t} (\|\xi - x^+\|^2 - \|\xi - x\|^2). \]

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

- At \( \xi = \xi^* \), with step size \( t = 1/L \),

\[
F(x^*) - F(x^+) \geq \frac{L}{2} (\|x^* - x^+\|^2 - \|x^* - x\|^2) + \frac{\nu}{2} \|x^* - x\|^2 \\
= \frac{L}{2} \|x^* - x^+\|^2 - \frac{L - \nu}{2} \|x^* - x\|^2.
\]
Linear convergence: Proof sketch

- Recall proof of prox-grad inequality:
  \[ F(\xi) - F(x^+) \geq \epsilon_f(x, \xi) + \frac{1}{2t} (\|\xi - x^+\|^2 - \|\xi - x\|^2). \]

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

- At \( \xi = \xi^* \), with step size \( t = 1/L \),
  \[
  F(x^*) - F(x^+) \geq \frac{L}{2} (\|x^* - x^+\|^2 - \|x^* - x\|^2) + \frac{\nu}{2} \|x^* - x\|^2
  = \frac{L}{2} \|x^* - x^+\|^2 - \frac{L - \nu}{2} \|x^* - x\|^2.
  \]

- Then, we use \( F(x^*) - F(x^+) \leq 0: \)
  \[
  \frac{L}{2} \|x^* - x^+\|^2 \leq \frac{L - \nu}{2} \|x^* - x\|^2 \iff \|x^* - x^+\|^2 \leq \left(1 - \frac{\nu}{L}\right) \|x^* - x\|^2.
  \]

- The rest of the proof follows by easy induction.
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
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 $x^{(k+1)}$ by taking a gradient step at point $y^{(k)}$, where $y^{(k)}$ is a well-chosen linear combination of previous 2 iterates, $x^{(k)}$ and $x^{(k-1)}$.
- Generalized to nonsmooth composite models by Beck & Teboulle [2009]. Method called FISTA for *fast iterative shrinkage-thresholding algorithm*, which describes the proximal gradient steps when $g(x) = \|x\|_1$. 
FISTA (here, with constant step sizes $t_k = \frac{1}{L}, \forall k$)

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

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

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

2. $\tau_{k+1} = \frac{1 + \sqrt{1 + 4 \tau_k^2}}{2}$

3. $y^{(k+1)} = x^{(k+1)} + \left( \frac{\tau_k - 1}{\tau_{k+1}} \right) (x^{(k+1)} - x^{(k)})$

\[\begin{align*}
\text{Initialization: } & y^{(0)} = x^{(0)} \in \text{int dom } f, \tau_0 = 1. \\
\text{For } k = 0, 1, 2, \ldots , & \\
1. & x^{(k+1)} = \text{prox}_{\frac{1}{L}g} (y^{(k)} - \frac{1}{L} \nabla f (y^{(k)})) \\
2. & \tau_{k+1} = \frac{1 + \sqrt{1 + 4 \tau_k^2}}{2} \\
3. & y^{(k+1)} = x^{(k+1)} + \left( \frac{\tau_k - 1}{\tau_{k+1}} \right) (x^{(k+1)} - x^{(k)})
\end{align*}\]
FISTA (here, with constant step sizes $t_k = \frac{1}{L}$, $\forall k$)

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

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

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

2. $\tau_{k+1} = \frac{1 + \sqrt{1 + 4\tau_k^2}}{2}$

3. $y^{(k+1)} = x^{(k+1)} + \left(\frac{\tau_k - 1}{\tau_{k+1}}\right)(x^{(k+1)} - x^{(k)})$

- In fact, $\tau_{k+1}$ 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]:
```python
# 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]:
```python
# generate data
In this example, $y = Ax_0 + \text{noise}$, 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.randn(m)
lbda = 0.05
```
Example: Lasso regression (2/4)

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

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

Compute Lipschitz constant

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

Proximal Gradient Method

```python
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

```python
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)

```python
In [18]: opt = \min(\min(\text{prox\_grad}), \min(\text{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 $\delta_k \leq \frac{LR^2}{2k}$ and $\delta_k \leq \frac{2LR^2}{(k+1)^2}$ are very pessimistic:

  After $k = 10^4$ iterations, assuming the exact value of $R = \|x^{(0)} - x^*\|$ is known,

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>$\delta_k$</th>
<th>upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proximal gradient</td>
<td>30.77</td>
<td>1421.41</td>
</tr>
<tr>
<td>FISTA</td>
<td>$1.29 \cdot 10^{-5}$</td>
<td>0.5684</td>
</tr>
</tbody>
</table>

- In practice, we can use much better duality bound on $\delta_k$ 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
Ω(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 \( f : \mathbb{R}^{2k+1} \rightarrow \mathbb{R} \) which is twice differentiable and \( L \)-smooth, such that for any sequence \( (x^{(i)})_{i \in \mathbb{N}} \) satisfying

\[
x^{(i+1)} \in x^{(0)} + \text{span}(\nabla f(x^{(0)}), \ldots, \nabla f(x^{(i)})), \quad \forall i \in \mathbb{N},
\]

it holds

\[
f(x^{(k)}) - f(x^*) \geq \frac{3L\|x^{(0)} - x^*\|^2}{32(k + 1)^2}.
\]
Ω(1/k^2) lower bound: Proof sketch

\[ f_k(x) := \frac{L}{4} \left( \frac{1}{2} x^T A x - e_1^T x \right), \]
where

\[ A = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ \vdots & \vdots & \vdots \\ -1 & 2 & -1 \end{pmatrix} \in S^k. \]

Assume (w.l.o.g.) that \( x^{(0)} = 0 \). We can show that
Ω(1/k^2) lower bound: Proof sketch

\[ f_k(x) := \frac{L}{4} \left( \frac{1}{2} x^T Ax - e_1^T x \right), \] where \( A = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ \vdots & \vdots & \vdots \\ -1 & 2 & -1 \end{pmatrix} \in \mathbb{S}^k. \]

Assume (w.l.o.g.) that \( x^{(0)} = 0 \). We can show that
- \( A \succeq 0 \) and \( \lambda_{\text{max}}(A) \leq 4 \), hence \( f_k \) is convex and \( L \)-smooth, \( \forall k \).
Ω(1/k^2) lower bound: Proof sketch

\[ f_k(x) := \frac{L}{4} \left( \frac{1}{2} x^T A x - e_1^T x \right), \]
where
\[
A = \begin{pmatrix}
2 & -1 & -1 \\
-1 & 2 & -1 \\
. & . & . \\
. & . & . \\
-1 & 2 & -1 \\

d\end{pmatrix} \in \mathbb{S}^k.
\]

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

- \( A \succeq 0 \) and \( \lambda_{\text{max}}(A) \leq 4 \), hence \( f_k \) is convex and \( L \)-smooth, \( \forall k \).
- \( f_k \) is minimized over \( \mathbb{R}^k \) at \( x^* = A^{-1} e_1 \), and

\[
f_k(x^*) = -\frac{L}{8} e_1^T A^{-1} e_1 = -\frac{L}{8} \left( 1 - \frac{1}{k + 1} \right).
\]
Ω(1/k²) lower bound: Proof sketch

\[ f_k(x) := \frac{L}{4} \left( \frac{1}{2} x^T A x - e_1^T x \right) \]
where \( A = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ & & \ddots & \ddots & \ddots \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix} \in \mathbb{S}^k. \)

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

- \( A \succeq 0 \) and \( \lambda_{\text{max}}(A) \leq 4, \) hence \( f_k \) is convex and \( L \)-smooth, \( \forall k. \)
- \( f_k \) is minimized over \( \mathbb{R}^k \) at \( x^* = A^{-1} e_1, \) and
  \[ f_k(x^*) = -\frac{L}{8} e_1^T A^{-1} e_1 = -\frac{L}{8} (1 - \frac{1}{k+1}). \]
- Let \( f = f_{2k+1}. \) A simple induction shows that for all \( i < 2k+1, \)
  \[ \text{span}(\nabla f(x^{(0)}), \ldots, \nabla f(x^{(i)})) \subseteq \text{span}(e_1, \ldots, e_{i+1}). \]
$\Omega(1/k^2)$ lower bound: Proof sketch

$$f_k(x) := \frac{L}{4} \left( \frac{1}{2} x^T Ax - e_1^T x \right), \text{ where } A = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ \vdots & \vdots & \vdots \\ -1 & 2 & -1 \end{pmatrix} \in \mathbb{S}^k.$$

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

- $A \succeq 0$ and $\lambda_{\text{max}}(A) \leq 4$, hence $f_k$ is convex and $L$-smooth, $\forall k$.
- $f_k$ is minimized over $\mathbb{R}^k$ at $x^* = A^{-1}e_1$, and
  $$f_k(x^*) = -\frac{L}{8} e_1^T A^{-1} e_1 = -\frac{L}{8} (1 - \frac{1}{k+1}).$$

- Let $f = f_{2k+1}$. A simple induction shows that for all $i < 2k+1$,
  $$\text{span}(\nabla f(x^{(0)}), \ldots, \nabla f(x^{(i)})) \subseteq \text{span}(e_1, \ldots, e_{i+1}).$$

- So, $f(x^{(k)}) \geq \inf_z f_k(z) = -\frac{L}{8} (1 - \frac{1}{k+1})$ and $f(x^*) = -\frac{L}{8} (1 - \frac{1}{2k+2})$. 
\( \Omega(1/k^2) \) lower bound: Proof sketch

\[
f_k(x) := \frac{L}{4} \left( \frac{1}{2} x^T Ax - e_1^T x \right), \text{ where } A = \begin{pmatrix} \frac{2}{1} & \frac{-1}{2} & \frac{-1}{2} \\ \frac{-1}{2} & \frac{2}{1} & \frac{-1}{2} \\ \vdots & \vdots & \vdots \\ \frac{-1}{2} & \frac{-1}{2} & \frac{2}{1} \end{pmatrix} \in \mathbb{S}^k.
\]

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

- \( A \succeq 0 \) and \( \lambda_{\text{max}}(A) \leq 4 \), hence \( f_k \) is convex and \( L \)-smooth, \( \forall k \).
- \( f_k \) is minimized over \( \mathbb{R}^k \) at \( x^* = A^{-1} e_1 \), and
  \[
f_k(x^*) = -\frac{L}{8} e_1^T A^{-1} e_1 = -\frac{L}{8} (1 - \frac{1}{k+1}).
\]

- Let \( f = f_{2k+1} \). A simple induction shows that for all \( i < 2k + 1 \),
  \[
  \text{span}(\nabla f(x^{(0)}), \ldots, \nabla f(x^{(i)})) \subseteq \text{span}(e_1, \ldots, e_{i+1}).
  \]

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

- Finally, we can bound \( \|x^*\|^2 \leq \frac{2}{3} (k + 1) \). Putting all together yields the desired bound.