# 6th Traditional Summer School for Young Researchers: "Control. Information. Optimization"

Giuseppe Carlo Calafiore

Dipartimento di Automatica e Informatica
Politecnico di Torino – ITALY

Moscow, June 22-29

# LECTURE 3

# The Proximal Gradient Method

# Outline

# Introduction

- In this lecture we discuss a first-order technique for solving constrained convex optimization problems of the form

$$\begin{aligned} p^* = \quad & \min_{x \in \mathbb{R}^n} f_0(x) \\ \text{s.t.} \quad & x \in \mathcal{X}, \end{aligned}$$

  where $f_0$ is a convex and differentiable function, and $\mathcal{X}$ is a convex set of simple structure. (we shall soon define what we mean by "simple").

- This method follows as a special case of a more general family of techniques used for solving a class of optimization problems with mixed differentiable plus non-differentiable objective, based on the concept of *proximal mapping*.

# Proximal mapping and projections

- Given a closed convex function $h : \mathbb{R}^n \to \mathbb{R}$ (not necessarily differentiable), we define the proximal mapping of $h$ as follows:

$$\text{prox}_h(x) = \arg\min_z \left( h(z) + \frac{1}{2}\|z - x\|_2^2 \right).$$

- Since $h(z)$ is convex and the additional term $\|z - x\|_2^2$ is strongly convex, then for each $x$ the function $h(z) + 0.5\|z - x\|_2^2$ is also strongly convex.
- Moreover, convexity of $h$ implies that $h(z) \geq h(x) + \eta_x^\top (z - x)$, for all $x$ in the interior of $\text{dom}\, h$, where $\eta_x$ is a subgradient of $h$ at $x$. Hence

$$h(z) + \frac{1}{2}\|z - x\|_2^2 \geq h(x) + \eta_x^\top (z - x) + \frac{1}{2}\|z - x\|_2^2,$$

  which implies that the function on the left of this inequality is bounded below.

- This property, together with strong convexity, guarantees that the global minimizer $\text{prox}_h(x)$ is well defined, since it exists and it is unique.

# Proximal mapping and projections

- An interesting special case arises when $h(z)$ is the indicator function of a closed convex set $\mathcal{X}$, i.e.,

$$h(z) = I_{\mathcal{X}}(z) \doteq \begin{cases} 0 & \text{if } z \in \mathcal{X}, \\ +\infty & \text{otherwise.} \end{cases}$$

- In this case, we have

$$\text{prox}_{I_{\mathcal{X}}}(x) = \arg\min_{z \in \mathcal{X}} \frac{1}{2}\|z - x\|_2^2, \tag{1}$$

  hence $\text{prox}_{I_{\mathcal{X}}}(x) = [x]_{\mathcal{X}}$ is the Euclidean projection of $x$ onto $\mathcal{X}$.

- We next refer to as *simple* those functions $h$ for which it is easy to compute the proximal mapping. Accordingly, we denote as *simple* a convex set $\mathcal{X}$ onto which it is computationally easy to determine a projection.

# Constrained problems in unconstrained format

- We observe that the constrained minimization problem (1) can be rewritten into an unconstrained form

$$\min_x \; f_0(x) + I_{\mathcal{X}}(x), \tag{2}$$

where the indicator function $I_{\mathcal{X}}(x)$ acts as a non-differentiable barrier for the feasible set $\mathcal{X}$.

- In the next slides we discuss an algorithm for solving a more general class of problems of the form

$$\min_x \; f_0(x) + h(x), \tag{3}$$

where $f_0$ is convex and differentiable, and $h(x)$ is convex and "simple." Problem (3) clearly includes (2), for $h(x) = I_{\mathcal{X}}(x)$.

# Proximal gradient method

- We address the solution of the problem

$$\min_x \ f(x), \tag{4}$$

where

$$f(x) \doteq f_0(x) + h(x),$$

via a modification of the gradient algorithm, adapted to the current situation where $h(x)$ may not be differentiable (hence its gradient may not exist).

- Given a current point $x_k$, the approach is to first perform a standard gradient step (using the gradient of $f_0$ only), and then compute the new point $x_{k+1}$ via the proximal map of $h$.
- In formulas, we take

$$x_{k+1} = \mathrm{prox}_{s_k h}(x_k - s_k \nabla f_0(x_k)),$$

where $s_k > 0$ is a stepsize.

- We next show how this update can be interpreted in terms of a "modified" gradient step.

# Proximal gradient method

- By the definition of proximal map, we have

$$
\begin{aligned}
x_{k+1} &= \mathrm{prox}_{s_k h}(x_k - s_k \nabla f_0(x_k)) \\
&= \arg\min_z \left( s_k h(z) + \frac{1}{2}\|z - x_k + s_k \nabla f_0(x_k)\|_2^2 \right) \\
&\quad \text{[dividing by } s_k \text{ does not change the minimizer]} \\
&= \arg\min_z \left( h(z) + \frac{1}{2s_k}\|(z - x_k) + s_k \nabla f_0(x_k)\|_2^2 \right) \\
&= \arg\min_z \left( h(z) + \frac{1}{2s_k}\|z - x_k\|_2^2 + \nabla f_0(x_k)^\top (z - x_k) + \right. \\
&\qquad \left. + \frac{s_k}{2}\|\nabla f_0(x_k)\|_2^2 \right) \\
&\quad \text{[adding the constant term } f_0(x_k) - \tfrac{s_k}{2}\|\nabla f_0(x_k)\|_2^2 \\
&\quad \text{does not change the minimizer]} \\
&= \arg\min_z \left( h(z) + f_0(x_k) + \nabla f_0(x_k)^\top (z - x_k) + \frac{1}{2s_k}\|z - x_k\|_2^2 \right).
\end{aligned}
$$

# Proximal gradient method

- The interpretation of this last formulation is that the updated point $x_{k+1}$ is the minimizer of $h(z)$ plus a local quadratic approximation of $f_0(z)$ at $x_k$, that is $x_{k+1} = \arg\min_z \psi_k(z)$, where

$$\begin{aligned} \psi_k(z) &\doteq h(z) + q_k(z), \\ q_k(z) &\doteq f_0(x_k) + \nabla f_0(x_k)^\top (z - x_k) + \frac{1}{2s_k}\|z - x_k\|_2^2. \end{aligned} \tag{5}$$

- Let us define a vector $g_s(x)$ as follows:

$$g_s(x) \doteq \frac{1}{s}(x - \mathrm{prox}_{sh}(x - s\nabla f_0(x)),$$

and also set, for notational simplicity,

$$g_k \doteq g_{s_k}(x_k) = \frac{1}{s_k}(x_k - x_{k+1}).$$

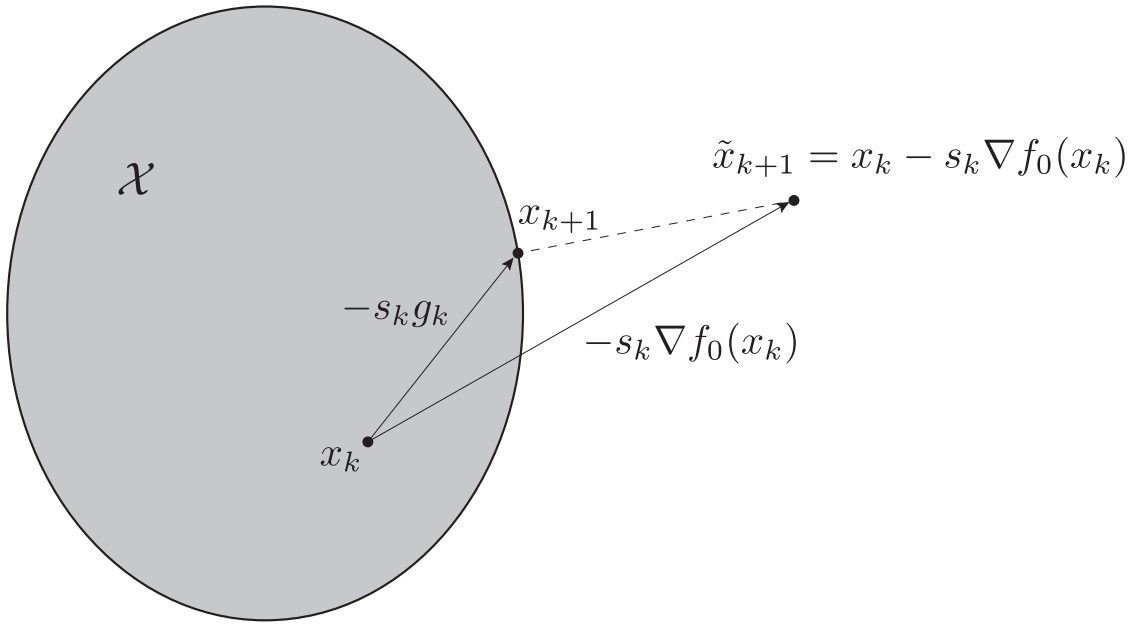
- With the above notation, we can formally write our algorithm as

$$x_{k+1} = x_k - s_k g_k, \tag{6}$$

where $g_k$ has the role of a "pseudo" gradient, and it is called the *gradient map* of $f_0$ on $h$ at $x_k$.

# Proximal gradient method

- Indeed, $g_k$ inherits some of the key properties of a standard gradient.
- For instance, it can be proved that the optimality condition for (4) is $g_s(x) = 0$.
- Also, if $h = 0$, then $g_k = \nabla f_0(x_k)$, hence $g_k$ is simply the gradient of $f_0$ at $x_k$.
- If instead $h = I_{\mathcal{X}}$, then the geometrical meaning of the gradient map is illustrated in the figure below: in this case $x_{k+1}$ is the Euclidean projection of $\tilde{x}_{k+1}$ onto $\mathcal{X}$.



# Proximal gradient method

A version of the proximal gradient algorithm, with constant stepsizes, is formally stated next.

**Require:** $f_0$ convex, differentiable, bounded below, with Lipschitz continuous gradient (Lipschitz constant $L$); $h$ convex and closed, $x_0 \in \operatorname{dom} f_0$, $\epsilon > 0$

1: Set $k = 0$, $s = 1/L$
2: Update: $x_{k+1} = \operatorname{prox}_{sh}(x_k - s\nabla f_0(x_k))$
3: If accuracy $\epsilon$ is attained (see, e.g., (13)), then exit and return $x_k$, else let $k \leftarrow k + 1$ and go to 2.

# Convergence of the proximal gradient algorithm

- We shall next prove convergence under the hypothesis that $f_0$ is strongly convex, with Lipschitz continuous gradient on $\mathrm{dom}\, f_0$. We start by observing that

$$\nabla q_k(z) = \nabla f_0(x_k) + \frac{1}{s_k}(z - x_k),$$

with $q_k$ defined in (5), hence

$$\nabla q_k(x_{k+1}) = \nabla f_0(x_k) - g_k.$$

- A point $x_k$ is optimal for problem (4), i.e., it minimizes $f = f_0 + h$, if and only if $g_k = 0$ (we skip the proof of this fact).
- Since $x_{k+1}$ minimizes $\psi_k(z)$, from the optimality conditions we have that

$$0 \in \partial h(x_{k+1}) + \nabla q_k(x_{k+1}) = \partial h(x_{k+1}) + \nabla f_0(x_k) - g_k,$$

that is

$$g_k \in \partial h(x_{k+1}) + \nabla f_0(x_k). \tag{7}$$

# Convergence of the proximal gradient algorithm

- Note that (7) means that there exists a subgradient $\eta_{k+1} \in \partial h(x_{k+1})$, such that

$$\nabla f_0(x_k) = g_k - \eta_{k+1},$$

where, by definition of a subgradient, it holds that

$$h(z) \geq h(x_{k+1}) + \eta_{k+1}^\top (z - x_{k+1}), \quad \forall z \in \operatorname{dom} h.$$

The last two relations will be useful soon.

- Now, the assumptions of strong convexity and Lipschitz continuous gradient on $f_0$ imply that there exists $m, L > 0$, $L \geq m$, such that

$$\begin{aligned} f_0(z) &\geq f_0(x_k) + \nabla f_0(x_k)^\top (z - x_k) + \frac{m}{2}\|z - x_k\|_2^2, \quad \forall z \in \operatorname{dom} f_0, \\ f_0(z) &\leq f_0(x_k) + \nabla f_0(x_k)^\top (z - x_k) + \frac{L}{2}\|z - x_k\|_2^2, \quad \forall z \in \operatorname{dom} f_0. \end{aligned}$$

- The second of these inequalities, evaluated at $z = x_{k+1}$, and for stepsizes such that $1/s_k \geq L$, yields

$$f_0(x_{k+1}) \leq q_k(x_{k+1}).$$

# Convergence of the proximal gradient algorithm

## A key inequality

- From the first of the previous inequalities (adding $h(z)$ on both sides) we have instead that, for all $z \in \operatorname{dom} f_0$,

$$f(z) - \frac{m}{2}\|z - x_k\|_2^2 \quad \geq \quad f(x_{k+1}) + \frac{s_k}{2}\|g_k\|_2^2 + g_k^\top (z - x_k). \qquad (8)$$

- (this inequality follows after a long and non-obvious series of steps that we omit.)

# Convergence of the proximal gradient algorithm

- Using this inequality for $z = x_k$, we obtain

$$f(x_{k+1}) \leq f(x_k) - \frac{s_k}{2}\|g_k\|_2^2,$$

which shows that the proximal gradient algorithm is a descent algorithm, and using again inequality (8) for $z = x^*$, where $x^*$ is the minimizer of $f$ we are seeking (hence $f(x_{k+1}) \geq f(x^*)$), we get

$$g_k^\top (x_k - x^*) \geq \frac{s_k}{2}\|g_k\|_2^2 + \frac{m}{2}\|x_k - x^*\|_2^2. \tag{9}$$

- Further, rewriting (8) as

$$f(z) \geq f(x_{k+1}) + \frac{s_k}{2}\|g_k\|_2^2 + g_k^\top (z - x_k) + \frac{m}{2}\|z - x_k\|_2^2, \quad \forall z \in \operatorname{dom} f_0, \tag{10}$$

and minimizing both sides over $z$ (note that the minimum of the expression on the right is attained at $z = x_k - (1/m)g_k$), we obtain

$$f(x_{k+1}) - f(x^*) \leq \frac{1}{2}\|g_k\|_2^2(1/m - s_k), \tag{11}$$

where $1/m - s_k \geq 0$, since this is implied by $L \geq m$ and $s_k \leq 1/L$.

# Convergence of the proximal gradient algorithm

- Also, evaluating (10) at $z = x^*$, we obtain

$$\begin{aligned} f(x_{k+1}) - f(x^*) &\leq g_k^\top (x_k - x^*) - \frac{s_k}{2}\|g_k\|_2^2 - \frac{m}{2}\|x_k - x^*\|_2^2 \\ &\leq g_k^\top (x_k - x^*) - \frac{s_k}{2}\|g_k\|_2^2 \\ &= \frac{1}{2s_k}\left(\|x_k - x^*\|_2^2 - \|x_k - x^* - s_k g_k\|_2^2\right) \\ &= \frac{1}{2s_k}\left(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2\right). \end{aligned} \tag{12}$$

- We next wrap up all these preliminaries. To derive our final result, let us consider, for simplicity, the case of constant stepsizes $s_k = s = 1/L$ (the proof can be adapted also for the case of stepsizes obtained via backtracking line search).

# Convergence of the proximal gradient algorithm

- Recalling (6), we have

$$\begin{aligned} \|x_{k+1} - x^*\|_2^2 &= \|(x_k - x^*) - s_k g_k\|_2^2 \\ &= \|x_k - x^*\|_2^2 + s_k^2 \|g_k\|_2^2 - 2 s_k g_k^\top (x_k - x^*) \\ \text{[using (9)]} &\leq (1 - m s_k) \|x_k - x^*\|_2^2 \\ \text{[for } s_k = 1/L\text{]} &= \left(1 - \frac{m}{L}\right) \|x_k - x^*\|_2^2, \end{aligned}$$

  whence

$$\|x_k - x^*\|_2^2 \leq \left(1 - \frac{m}{L}\right)^k \|x_0 - x^*\|_2^2,$$

  which shows that the proximal gradient algorithm converges to $x^*$ at a linear rate.

- Also, if $m, L$ are known, then (11) provides a stopping criterion based on checking the norm of $g_k$, since, for $\epsilon \geq 0$,

$$\|g_k\|_2^2 \leq 2\epsilon \frac{mL}{L - m} \quad \Rightarrow \quad f(x_{k+1}) - f(x^*) \leq \epsilon. \tag{13}$$

# Convergence of the proximal gradient algorithm

- Further, adding inequalities (12), we get

$$\begin{aligned}\sum_{i=1}^{k} f(x_i) - f(x^*) &\leq \frac{1}{2s}\sum_{i=1}^{k}\left(\|x_{i-1}-x^*\|_2^2 - \|x_i - x^*\|_2^2\right)\\ &= \frac{L}{2}\left(\|x_0-x^*\|_2^2 - \|x_k-x^*\|_2^2\right)\\ &\leq \frac{L}{2}\|x_0-x^*\|_2^2.\end{aligned}$$

- Since $f(x_i)$ is non-increasing with $i$, the last value $f(x_k)$ is no larger than the average of the previous values, that is

$$f(x_k) - f(x^*) \leq \frac{1}{k}\sum_{i=1}^{k}(f(x_i) - f(x^*)) \leq \frac{L}{2k}\|x_0 - x^*\|_2^2,$$

which shows that $f(x_k) \to f(x^*)$ at rate $1/k$.

# Convergence of the proximal gradient algorithm

Our findings are summarized in the next theorem.

## Theorem 1

*For the proximal gradient algorithm it holds that*

$$f(x_k) - f(x^*) \leq \frac{L}{2k}\|x_0 - x^*\|_2^2.$$

*Moreover, under the additional hypothesis that $f_0$ is strongly convex (with strong convexity constant $m$), it also holds that*

$$\begin{aligned} \|x_k - x^*\|_2^2 &\leq \left(1 - \frac{m}{L}\right)^k \|x_0 - x^*\|_2^2, \\ f(x_{k+1}) - f(x^*) &\leq \frac{1}{2}\|g_k\|_2^2(1/m - 1/L). \end{aligned}$$

# Computing proximal maps and projections

- We next discuss several relevant cases of functions $h$ for which the proximal maps are "easy" to compute.
- We recall that if $h$ is the indicator function of a closed convex set $\mathcal{X}$, then the proximal map is just the Euclidean projection onto $\mathcal{X}$ hence, in this case, the proximal gradient algorithm solves the constrained optimization problem

$$\begin{aligned} p^* = \quad & \min_{x \in \mathbb{R}^n} \ f_0(x) \\ \text{s.t.} \quad & x \in \mathcal{X}. \end{aligned}$$

# Computing proximal maps and projections

## Projection onto a half-space.

- Let $\mathcal{X}$ be a half-space

$$\mathcal{X} = \{x : a^\top x \leq b\}, \quad a \neq 0.$$

- Then, for given $x \in \mathbb{R}^n$, we have

$$\text{prox}_{I_\mathcal{X}}(x) = \arg\min_{z \in \mathcal{X}} \|z - x\|_2^2 = [x]_\mathcal{X},$$

and the projection $[x]_\mathcal{X}$ is $x$, if $x \in \mathcal{X}$, or it is equal to the projection of $x$ onto the hyperplane $\{x : a^\top x = b\}$, if $x \notin \mathcal{X}$.

- This latter projection is given by

$$[x]_\mathcal{X} = \begin{cases} x & \text{if } a^\top x \leq b, \\ x + \frac{b - a^\top x}{\|a\|_2^2} a & \text{if } a^\top x > b. \end{cases}$$

# Computing proximal maps and projections

## Projection onto the positive orthant.

- Let
$$\mathcal{X} = \mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x \geq 0\}.$$
- Then, for given $x \in \mathbb{R}^n$, we have
$$[x]_{\mathcal{X}} = \arg\min_{z \geq 0} \|z - x\|_2^2 = \arg\min_{z_i \geq 0} \sum_{i=1}^{n} (z_i - x_i)^2,$$
where we see that the optimal $z$ should have components $z_i = x_i$, if $x_i \geq 0$, or $z_i = 0$ otherwise, hence
$$[x]_{\mathcal{X}} = [x]_+ = \max(0, x),$$
where the max is here intended element-wise.

# Computing proximal maps and projections

## Projection onto the standard simplex.

- Let $\mathcal{X}$ be the standard (probability) simplex

$$\mathcal{X} = \{x \in \mathbb{R}^n : x \geq 0, \mathbf{1}^\top x = 1\}.$$

- Computing the projection $[x]_{\mathcal{X}}$ amounts to solving $\min_{z \in \mathcal{X}} \frac{1}{2}\|z - x\|_2^2$.
- Considering the (partial) Lagrangian for this problem, we have

$$\mathcal{L}(z, \nu) = \frac{1}{2}\|z - x\|_2^2 + \nu(\mathbf{1}^\top z - 1)$$

and the dual function

$$\begin{aligned} g(\nu) &= \min_{z \geq 0} \mathcal{L}(z, \nu) = \min_{z \geq 0} \frac{1}{2}\|z - x\|_2^2 + \nu(\mathbf{1}^\top z - 1) \\ &= \min_{z \geq 0} \sum_{i=1}^n \left( \frac{1}{2}(z_i - x_i)_2^2 + \nu z_i \right) - \nu \\ &= \sum_{i=1}^n \min_{z_i \geq 0} \left( \frac{1}{2}(z_i - x_i)_2^2 + \nu z_i \right) - \nu. \end{aligned}$$

# Computing proximal maps and projections

## Projection onto the standard simplex.

- The dual function is *separable*: optimal solution $z$ is obtained by finding the optimal values of the individual components, which are obtained by solving a simple one-dimensional minimization: $z_i^*(\nu) = \arg\min_{z_i \geq 0} \frac{1}{2}(z_i - x_i)_2^2 + \nu z_i$.
- The function to be minimized here is a convex parabola, having its vertex at $v_i = x_i - \nu$. The minimizer $z_i^*(\nu)$ is thus the vertex, if $v_i \geq 0$, or it is zero otherwise. That is $z^*(\nu) = \max(x - \nu \mathbf{1}, 0)$.
- The optimal value $\nu^*$ of the dual variable $\nu$ should be then obtained by maximizing $g(\nu)$ with respect to $\nu$. However, there exists only one value of $\nu$ which makes $z^*(\nu)$ belong to the simplex (i.e., primal feasible: $\sum_i z_i^*(\nu) = 1$), hence this must be the optimal value of the dual variable.
- In summary

$$[x]_{\mathcal{X}} = z^*(\nu^*) = \max(x - \nu^* \mathbf{1}, 0),$$

where $\nu^*$ is the solution of the scalar equation

$$\sum_{i=1}^{n} \max(x_i - \nu, 0) = 1.$$

# Computing proximal maps and projections

## Projection onto the Euclidean ball.

- Let $\mathcal{X}$ be the unit Euclidean ball

$$\mathcal{X} = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}.$$

- Then, it is straightforward to verify that the projection of $x$ onto $\mathcal{X}$ is

$$[x]_{\mathcal{X}} = \begin{cases} x & \text{if } \|x\|_2 \leq 1, \\ \frac{x}{\|x\|_2} & \text{if } \|x\|_2 > 1. \end{cases}$$

# Computing proximal maps and projections

## Projection onto the $\ell_1$-norm ball.

- Let $\mathcal{X}$ be the unit $\ell_1$ ball

$$\mathcal{X} = \{x \in \mathbb{R}^n : \|x\|_1 \leq 1\}.$$

- For a given $x \in \mathbb{R}^n$, computing the projection $[x]_{\mathcal{X}}$ amounts to solving

$$\min_{\|z\|_1 \leq 1} \frac{1}{2}\|z - x\|_2^2. \tag{14}$$

- The Lagrangian of this problem is

$$\mathcal{L}(z, \lambda) = \frac{1}{2}\|z - x\|_2^2 + \lambda(\|z\|_1 - 1),$$

hence the dual function is

$$\begin{aligned} q(\lambda) &= \min_z \mathcal{L}(z, \lambda) = \min_z \frac{1}{2}\|z - x\|_2^2 + \lambda(\|z\|_1 - 1) \qquad (15) \\ &= \sum_{i=1}^n \min_{z_i} \left( \frac{1}{2}(z_i - x_i)^2 + \lambda |z_i| \right) - \lambda. \end{aligned}$$

# Computing proximal maps and projections

## Projection onto the $\ell_1$-norm ball.

- We then see that the values $z_i^*(\lambda)$ that minimize the above function are found by solving the following univariate minimizations:

$$z_i^*(\lambda) = \arg\min_{z_i} \varphi(z_i, \lambda), \quad \varphi(z_i, \lambda) \doteq \frac{1}{2}(z_i - x_i)^2 + \lambda|z_i|, \quad i = 1, \dots, n.$$

- To solve this problem, we use the identity $|z_i| = \max_{|\varrho_i| \leq 1} \varrho_i z_i$, and write

$$\begin{aligned}
\min_{z_i} \frac{1}{2}(z_i - x_i)^2 + \lambda|z_i| &= \min_{z_i} \left( \frac{1}{2}(z_i - x_i)^2 + \lambda \max_{|\varrho_i| \leq 1} \varrho_i z_i \right) \\
&= \min_{z_i} \max_{|\varrho_i| \leq 1} \frac{1}{2}(z_i - x_i)^2 + \lambda \varrho_i z_i \\
&= \max_{|\varrho_i| \leq 1} \min_{z_i} \frac{1}{2}(z_i - x_i)^2 + \lambda \varrho_i z_i.
\end{aligned}$$

- The inner minimization (w.r.t. $z_i$) is readily solved by setting the derivative to zero, obtaining $z_i^*(\lambda) = x_i - \lambda\varrho_i$, which, substituted back, yields

$$\min_{z_i} \frac{1}{2}(z_i - x_i)^2 + \lambda \varrho_i z_i = \lambda \left( \varrho_i x_i - \frac{1}{2}\lambda \varrho_i^2 \right).$$

# Computing proximal maps and projections

## Projection onto the $\ell_1$-norm ball.

- Continuing the previous chain of equalities, we thus have that

$$\min_{z_i} \frac{1}{2}(z_i - x_i)^2 + \lambda|z_i| = \lambda \max_{|\varrho_i|\leq 1} \left( \varrho_i x_i - \frac{1}{2}\lambda \varrho_i^2 \right).$$

- The function to be maximized here (w.r.t. $\varrho_i$) is a concave parabola, having its vertex at $v_i = x_i/\lambda$ (we let here $\lambda > 0$, since for $\lambda = 0$ the dual function is trivially zero). Hence, if $|v_i| \leq 1$ the maximum is attained at $\varrho_i^* = v_i = x_i/\lambda$. Otherwise, the maximum is attained at one of the extremes of the feasible interval $\varrho_i \in [-1, 1]$ and, in particular, at $\varrho_i^* = 1$ if $x_i \geq 0$, and at $\varrho_i^* = -1$ if $x_i < 0$. Therefore,

$$\varrho_i^* = \begin{cases} x_i/\lambda & \text{if } |x_i| \leq \lambda, \\ \operatorname{sgn}(x_i) & \text{otherwise.} \end{cases}$$

- Correspondingly, the minimizer $z_i^*(\lambda)$ of $\varphi(z_i, \lambda)$ is

$$z_i^*(\lambda) = x_i - \lambda \varrho_i^* = \begin{cases} 0 & \text{if } |x_i| \leq \lambda, \\ x_i - \lambda \operatorname{sgn}(x_i) & \text{otherwise.} \end{cases}$$

# Computing proximal maps and projections

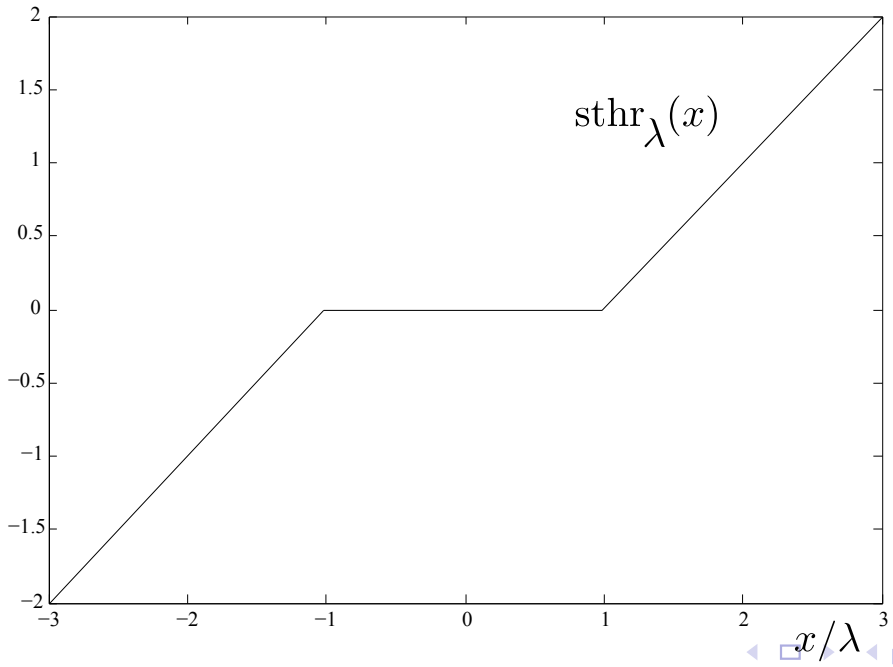
## Projection onto the $\ell_1$-norm ball.

- This can be more compactly written as

$$z_i^*(\lambda) = \mathrm{sgn}(x_i)[|x_i| - \lambda]_+ \doteq \mathrm{sthr}_\lambda(x_i), \quad i = 1, \ldots, n$$

  where $[\cdot]_+$ denotes the projection onto the positive orthant (positive part of the argument).

- The function $\mathrm{sthr}_\lambda$ is known as the *soft threshold* function, or *shrinkage operator*.



# Computing proximal maps and projections

## Projection onto the $\ell_1$-norm ball.

- Now, since strong duality holds for problem (14), and the solution to this problem is unique, the optimal primal variable $[x]_{\mathcal{X}}$ coincides with $z_i^*(\lambda^*)$, where $\lambda^* \geq 0$ is the value of the dual variable that maximizes $q(\lambda)$.
- It can be proved that the optimal $\lambda$ is the solution of the scalar equation

$$\sum_{i=1}^{n} \max(|x_i| - \lambda, 0) = 1.$$

- Once $\lambda^*$ is found, the projection we seek is given by

$$[x]_{\mathcal{X}} = \mathrm{sgn}(x_i)[|x_i| - \lambda^*]_+.$$

# Computing proximal maps and projections

## Projection onto the positive semidefinite cone.

- Consider the cone of positive semidefinite matrices

$$\mathcal{X} = \{X \in \mathbb{S}^n : X \succeq 0\} = \mathbb{S}^n_+.$$

- Given a matrix $X \in \mathbb{S}^n$ we want to compute its projection onto $\mathcal{X}$. Since we are working in a matrix space, we shall define projections according to the Frobenius norm, that is

$$[X]_{\mathcal{X}} = \arg \min_{Z \in \mathcal{X}} \|Z - X\|_{\mathrm{F}}^2.$$

- Let now $X = U\Lambda U^\top$ be a spectral factorization for $X$, where $U$ is an orthogonal matrix, and $\Lambda$ is diagonal, containing the eigenvalues of $X$ on the diagonal. Since the Frobenius norm is unitarily invariant, we have that

$$\begin{aligned} \|Z - X\|_{\mathrm{F}}^2 &= \|Z - U\Lambda U^\top\|_{\mathrm{F}}^2 = \|U(U^\top Z U - \Lambda)U^\top\|_{\mathrm{F}}^2 \\ &= \|U^\top Z U - \Lambda\|_{\mathrm{F}}^2 = \|\tilde{Z} - \Lambda\|_{\mathrm{F}}^2, \end{aligned}$$

where we defined $\tilde{Z} \doteq U^\top Z U$.

# Computing proximal maps and projections

Projection onto the positive semidefinite cone.

- Since $\Lambda$ is diagonal it is easy to see that the $\tilde{Z}$ that minimizes $\|\tilde{Z} - \Lambda\|_{\mathrm{F}}^2$ is also diagonal, and

$$\tilde{Z}^* = \operatorname{diag}\left([\lambda_1]_+, \ldots, [\lambda_n]_+\right) = [\Lambda]_+,$$

  whence $Z^* = U\tilde{Z}^*U^\top$.

- In summary, the projection of $X = U\Lambda U^\top$ onto the positive semidefinite cone is given by

$$[X]_{\mathbb{S}_+^n} = U[\Lambda]_+U^\top.$$

# Proximal map of $\ell_1$ regularization.

- In many problems of practical relevance the function $h$ is a scalar multiple of the $\ell_1$ norm of $x$.
- For instance, in the $\ell_1$-regularized least-squares problem (also known as the LASSO), we consider

$$\min_x \frac{1}{\gamma}\|Ax-b\|_2^2 + \|x\|_1, \tag{16}$$

which is of the form (4), with $f_0(x) = (1/\gamma)\|Ax-b\|_2^2$ strongly convex (assuming $A$ is full rank), and $h(x) = \|x\|_1$ convex but non-differentiable.

- This class of problems is thus solvable by means of the proximal gradient algorithm. To this end, we need to be able to efficiently compute the proximal map of $sh$, where $s \geq 0$ is a scalar (the stepsize), namely

$$\mathrm{prox}_{sh}(x) = \arg\min_z \; s\|x\|_1 + \frac{1}{2}\|z-x\|_2^2.$$

- We already showed that the solution is given by the soft threshold function:

$$\mathrm{prox}_{sh}(x) = \mathrm{sthr}_s(x),$$

where the $i$-th component of the vector $\mathrm{sthr}_s(x)$ is $\mathrm{sgn}(x_i)[|x_i| - s]_+$.

# Proximal gradient algorithm for the LASSO.

- We can specify the proximal gradient algorithm in the case of the LASSO problem in (16).
- Notice that we have in this case

$$\begin{aligned}\nabla f_0(x) &= \frac{2}{\gamma}\left(A^\top Ax - A^\top b\right),\\ \nabla^2 f_0(x) &= \frac{2}{\gamma}(A^\top A),\end{aligned}$$

from which it follows that the strong convexity constant for $f_0$ is

$$m = \frac{2}{\gamma}\sigma_{\min}(A^\top A). \tag{17}$$

- Further, we have that

$$\|\nabla f_0(x) - \nabla f_0(y)\|_2 = \frac{2}{\gamma}\|A^\top A(x-y)\|_2 \leq \frac{2}{\gamma}\sigma_{\max}(A^\top A)\|x-y\|_2,$$

from which we obtain a global Lipschitz constant for the gradient:

$$L = \frac{2}{\gamma}\sigma_{\max}(A^\top A). \tag{18}$$

# Proximal gradient algorithm for the LASSO.

## ISTA (iterative shrinkage–thresholding algorithm)

**Require:** $\epsilon > 0$, $x_0$, $A$ full rank.

1: Compute $m$, $L$ according to (17), (18)
2: Set $k = 0$, $s = 1/L$
3: Compute gradient $\nabla f_0(x_k) = (2/\gamma)(A^\top A x_k - A^\top b)$
4: Update: $x_{k+1} = \mathrm{sthr}_s(x_k - s\nabla f_0(x_k))$
5: Compute $\|g_k\|_2 = \|x_k - x_{k+1}\|_2/s$
6: If $\|g_k\|_2^2 \leq 2\epsilon mL/(L-m)$, then return $x = x_{k+1}$ and exit, else let $k \leftarrow k+1$ and go to 3.

# Proximal gradient algorithm for the LASSO.

- In recent years there has been a tremendous activity in theory, algorithms and applications of $\ell_1$ regularization and LASSO-related problems, especially in the context of the "compressive sensing" field. More sophisticated techniques thus exist for solving the LASSO and related problems.
- The key essential advance provided by these techniques consists of "accelerating" the basic proximal gradient scheme, so as to reach convergence rates of the order of $1/k^2$ (recall that the basic proximal gradient has convergence rate of the order of $1/k$ on objective value).
- We next present (without proof) one of these "fast" methods.

# Fast proximal gradient algorithm

## FISTA (fast iterative shrinkage–thresholding algorithm)

**Require:** $x_0$, a Lipschitz constant $L$ for $\nabla f_0$

1: Set $k = 1$, $s = 1/L$, $y_1 = x_0$, $t_1 = 1$
2: Update: $x_k = \mathrm{prox}_{sh}(y_k - s\nabla f_0(y_k))$
3: Update: $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
4: Update: $y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})$
5: If $\|x_k - x_{k-1}\|_2 \leq \epsilon$, then return $x = x_k$ and exit, else let $k \leftarrow k + 1$ and go to 2.

# FISTA

- When applied to the specific LASSO problem in (16), step 2 in this algorithm simply reduces to soft thresholding:

$$\text{prox}_{sh}(y_k - s\nabla f_0(y_k)) = \text{sthr}_s(y_k - s\nabla f_0(y_k)).$$

## Theorem 2 (A. Beck, M. Teboulle, 2009)

*For the sequence $x_k$, $k = 1, \ldots,$ generated FISTA it holds that*

$$f(x_k) - f(x^*) \leq \frac{2L}{(k+1)^2}\|x_0 - x^*\|_2^2,$$

*where $x^*$ is any optimal solution to problem (4).*