Chapter 5: First order methods

Edouard Pauwels

Statistics and optimization in high dimensions

M2RI, Toulouse 3 Paul Sabatier

$\mathbb{X} \in \mathbb{R}^{n \times d}$, $Y \in \mathbb{R}^n$ (random).

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \|\mathbb{X}\theta - Y\|_2^2$$
$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \|\mathbb{X}\theta - Y\|_2^2, \qquad \text{s.t.} \qquad \|\theta\|_1 \leq 1$$
$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \|\mathbb{X}\theta - Y\|_2^2, \qquad \text{s.t.} \qquad \|\theta\|_0 \leq k$$
$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \|\mathbb{X}\theta - Y\|_2^2 + \lambda\|\theta\|_0$$
$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \|\mathbb{X}\theta - Y\|_2^2 + \lambda\|\theta\|_1$$
$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \|\theta\|_1, \qquad \text{s.t.} \qquad \mathbb{X}\theta = Y.$$

- The first one can be computed in $O(n^2 d)$ operations.
- How about the other ones?
- How to deal with large values of n and d.

Where have we been so far.

- $\| \cdot \|_0$: hard to handle computationally.
- ℓ_1 norm estimators are solutions to conic programs.
- General purpose solvers (interior point methods).
- Iterative methods: at least d^3 per iteration.

Plan for today: if one cannot afford d^3. Introduction to first order methods and nonsmooth analysis.

- Analysis of gradient descent algorithm.
- Introduction to the notion of subgradient.
- Algorithm for nonsmooth optimization: subgradient and proximal gradient.
- Acceleration.

Sources are diverse, see the lecture notes.

Plan

1. Gradient descent algorithm
2. Nonsmooth analsysis
3. Subgradient descent
4. Composite optimization
5. Lower bounds and acceleration

Proposition

Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ be twice differentiable with compact sublevel sets. Consider the differential equation, for $x_0 \in \mathbb{R}^p$,

$$\dot{x}(t) = -\nabla f(x(t)) \tag{1}$$

$$x(0) = x_0. \tag{2}$$

Then, there exists a solution to the initial value problem defined for all $t > 0$.

- $\int_0^{+\infty} \|\nabla f(x(t))\|_2^2 dt < +\infty$ *and* $\lim_{t\to\infty} \|\nabla f(x(t))\| = 0$.
- *Any accumulation point $\bar{x}$ of the trajectory satisfies $\nabla f(\bar{x}) = 0$.*
- *If in addition f is convex, set $f^* = \inf_{x\in\mathbb{R}^p} f(x)$ and assume that it is attained at x^*, we have for any $t \in \mathbb{R}$, $t > 0$,*

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

And $x(t) \underset{t\to\infty}{\to} \bar{x}$ where $\bar{x}$ is a global minimizer of f.

Gradient descent and a descent lemma

Gradient algorithm: $f : \mathbb{R}^p \mapsto \mathbb{R}$, iteration cost of the order of p.

$$x_{k+1} = x_k - s_k \nabla f(x_k) \tag{3}$$

Lemma

Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ be continuously differentiable with L-Lipschitz gradient ($L > 0$), then for any $x, y \in \mathbb{R}^p$,

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|_2^2.$$

Gradient descent algorithm

Proposition

Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ *be continuously differentiable with L-Lipschitz gradient and such that* $\inf_{x \in \mathbb{R}^p} f(x) > -\infty$. *Consider the algorithm, for* $x_0 \in \mathbb{R}^p$ *and*

$$x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k). \tag{4}$$

Then

- $\lim_{k \to \infty} \|\nabla f(x_k)\| = 0$, *(any accumulation point* $\bar{x}$ *of the trajectory satisfies* $\nabla f(\bar{x}) = 0$*).*
- *If in addition* f *is convex, set* $f^* = \inf_{x \in \mathbb{R}^p} f(x)$ *and assume that it is attained at* x^**, we have for any* $k \in \mathbb{N}$, $k > 0$,

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

  Furthermore x_k *converges to* $\bar{x}$ *a global minimum of* f
- *If in addition* f *is* μ*-strongly convex, then we have for any* $k \in \mathbb{N}$

  $$f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right) (f(x_k) - f^*).$$

Plan

1. Gradient descent algorithm
2. Nonsmooth analsysis
3. Subgradient descent
4. Composite optimization
5. Lower bounds and acceleration

Nonsmooth analysis?

How to deal with ℓ_1 norm penalty? We need a generalization of the notion of gradient.

Lower semicontinuity: f denotes a lower semi-continuous convex function on $\mathbb{R}^p$. Lower semi-continuity: epigraph is closed:

$$\mathrm{epi}_f = \left\{(x, z) \in \mathbb{R}^{p+1},\, z \geq f(x)\right\}.$$

equivalently as for any $x \in \mathbb{R}^p$

$$\liminf_{y \to x} f(y) \geq f(x).$$

Domain: f is allowed to take value $+\infty$, we denote its domain by

$$\mathrm{dom}_f = \{x \in \mathbb{R}^p,\, f(x) < +\infty\},$$

which is a convex set.

For example:

$$\min_{\|\theta\|_1 \leq 1} \|\mathbb{X}\theta - y\|_2^2 = \min_{\theta \in \mathbb{R}^d} \|\mathbb{X}\theta - y\|_2^2 + \delta(x)$$

where $\delta(\theta) = 0$ if $\|\theta\|_1 \leq 1$, and $+\infty$ otherwise (indicator of the ℓ_1 unit ball).

Exercise

Show that a convex function is continuous on the interior of its domain.

Definition

For any $x \in \mathrm{dom}_f$, the subgradient of f denotes the set

$$\partial f(x) = \{v \in \mathbb{R}^p,\, f(y) \geq f(x) + \langle v, y - x \rangle \,,\, \forall y \in \mathbb{R}^p\} \,.$$

For $x \notin \mathrm{dom}_f$, $\partial f(x)$ is set to be empty.

We deduce from the definition the generalization of Fermat rule

Theorem

$x^* \in \arg\min_x f(x)$ *if and only if* $0 \in \partial f(x^*)$.

Exercise, indicator: Let C be compact convex and $\delta(x) = 0$ for $x \in C$ and $+\infty$ otherwise. Describe $\partial\delta$.

Properties of the subgradient

Proposition

For any $x \in \mathbb{R}^p$, $\partial f(x)$ is a closed convex set. Furthermore, at any $x \in \mathrm{int}(\mathrm{dom}_f)$, $\partial f(x)$ is non empty and bounded.

Exercise, sequential closedness: Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ be a convex function, show that ∂f is sequencialy closed in the sence that, for any $\bar{x}$

$$\left\{ v \in \mathbb{R}^p, \, \exists \left(x_k, v_k\right)_{k \in \mathbb{N}}, \, x_k \to \bar{x}, \, v_k \to v, \, v_k \in \partial f(x_k), \, f(x_k) \to f(\bar{x}) \right\} \subset \partial f(\bar{x})$$

Exercise, lipschitzness: Let $f : \mathbb{R}^p \mapsto \mathbb{R}$, show that f is L-Lipschitz if and only if $\sup_{x \in \mathbb{R}^p, \, v \in \partial f(x)} \|v\|_2 \leq L$.

Exercise, sum rule: Let f and g be convex. Show that $\partial(f+g)(x) \supset \partial f(x) + \partial g(x)$ for every x such that $\partial f(x)$ and $\partial g(x)$ are non empty. What do you think about the reverse inclusion?

Representation of convex functions

Theorem

Let f be convex and lower semicontinuous and finite at least at one point, then f is the supremum of all its affine minorants: for any $x \in \mathbb{R}^p$

$$f(x) = \sup_{r \in \mathbb{R}, v \in \mathbb{R}^p} r + v^T x \qquad \text{s.t.} \qquad f(y) \geq r + v^T y, \forall y \in \mathbb{R}^p.$$

Theorem

For any $x \in \mathrm{int}(\mathrm{dom}_f)$ *and any* $h \in \mathbb{R}^p$,

$$D_h f(x) = \sup_{v \in \partial f(x)} \langle v, h \rangle ,$$

where D_h *denotes the directional derivative of* f*,*

$$D_h f(x) = \lim_{t>0, t\to 0} \frac{f(x+th) - f(x)}{t}.$$

Consequence: if f is differentiable at $\bar{x}$ then $\partial f(\bar{x}) = \{\nabla f(\bar{x})\}$

Fenchel-Young's inequality

Definition

Given f convex, the Fenchel-Legendre transform of f is given as follows

$$f^*: z \mapsto \sup_{y \in \mathbb{R}^p} z^T y - f(y)$$

Theorem

For any f convex, f^* is convex and for any $x, z \in \mathbb{R}^p$

$$f(x) + f^*(z) \geq z^T x$$

and the preceeding inequality holds if and only if $z \in \partial f(x)$. This is called Fenchel-Young's inequality. Furthermore, f is lower semicontinuous if and only if $(f^*)^* = f$.

Exercise

Set $f\colon x \mapsto \|x\|_1$. Compute f^* and the subgradient of f.

Plan

1. Gradient descent algorithm
2. Nonsmooth analsysis
3. Subgradient descent
4. Composite optimization
5. Lower bounds and acceleration

Convergence analysis for subgradient descent

Proposition

Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ *be a convex function which attains its infimum and has full domain. Consider the algorithm, for* $x_0 \in \mathbb{R}^p$*, a sequence of positive numbers* $\alpha_k > 0$*,* $k \in \mathbb{N}$*, iterate*

$$x_{k+1} = x_k - \alpha_k v_k \tag{5}$$

$$v_k \in \partial f(x_k). \tag{6}$$

Then for any global minimizer x^**, setting,* $y_k = \sum_{i=0}^k \alpha_i x_i / \left(\sum_{i=0}^k \alpha_i\right)$

$$\min_{i=1,\ldots,k} f(x_k) - f^* \leq \frac{\|x_0 - x^*\|^2 + \sum_{i=0}^k \alpha_i^2 \|v_i\|_2^2}{2\sum_{i=0}^k \alpha_i}$$

$$f(y_k) - f^* \leq \frac{\|x_0 - x^*\|^2 + \sum_{i=0}^k \alpha_i^2 \|v_i\|_2^2}{2\sum_{i=0}^k \alpha_i}.$$

Corollary

If f is L-Lipschitz, we have the following convergence result for subgradient method.

- *If $\alpha_k = \alpha$ is constant, we have*

$$\min_{i=1,\ldots,k} f(x_k) - f^* \leq \frac{\|x_0 - x^*\|^2}{2(k+1)\alpha} + \frac{L^2\alpha}{2}.$$

- *In particular, choosing $\alpha_i = \frac{\|x_0 - x^*\|/L}{\sqrt{k+1}}$, we have*

$$\min_{i=1,\ldots,k} f(x_k) - f^* \leq \frac{\|x_0 - x^*\| L}{\sqrt{k+1}}.$$

- *Choosing $\alpha_k = \|x_0 - x^*\|/(L\sqrt{k})$ for all k, we obtain for all k*

$$\min_{i=1,\ldots,k} f(x_k) - f^* = O\left(\frac{\|x_0 - x^*\|_2 L(1 + \log(k))}{\sqrt{k}}\right).$$

Subgradient algorithm

- Very generic, applies to any convex function (all purpose tool).
- Requires computing subgradient (interior of the domain).
- The convergence rate is optimal among all Lipschitz function.
- Hard to tune (decreasing step size).
- Quite slow in practice.

Plan

Motivation

Lasso estimator

$$\hat{\theta}^{\ell_1} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \|\mathbb{X}\theta - Y\|^2 + \lambda \|\theta\|_1.$$

Composite structure: "Smooth + non smooth"

$$F = f + g$$

where f is smooth and g is convex, non smooth.

This additional structure can be leveraged. This model includes constrained optimization problems.

Proximity operator (Moreau)

Definition

Given a closed convex function, $g \colon \mathbb{R}^d \mapsto \mathbb{R}$, the proximity operator of g is defined as follows

$$\mathrm{prox}_f \colon z \mapsto \arg\min_{y \in \mathbb{R}^d} g(y) + \frac{1}{2}\|y - z\|_2^2 .$$

By strong convexity, the minimum is attained and is strict.

Note that we have $x = \mathrm{prox}_g(z)$ if and only if $z = \partial g(x) + x$ and the proximity operator is sometimes denoted $(\partial g + I)^{-1}$.

Indicator: Let C be compact convex and $\delta(x) = 0$ for $x \in C$ and $+\infty$ otherwise. What is prox_δ?

Extension of the descent Lemma

$$\min_{x \in \mathbb{R}^p} f(x) + g(x)$$

Lemma

Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ be convex continuously differentiable with L-Lipschitz gradient and g be convex lower semicontinuous. Fix any $x \in \mathbb{R}^p$ and set

$$y = \mathrm{prox}_{g/L}\left(x - \frac{1}{L}\nabla f(x)\right).$$

Then, for any $z \in \mathbb{R}^d$,

$$f(z) + g(z) + \frac{L}{2}\|x - z\|_2^2 \geq f(y) + g(y) + \frac{L}{2}\|y - z\|_2^2.$$

Proximal gradient algorithm

$$\min_{x \in \mathbb{R}^p} f(x) + g(x)$$

Proposition

Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ be convex continuously differentiable with L-Lipschitz gradient and g be convex lower semicontinuous such that $\rho = \inf_{x \in \mathbb{R}^p} f(x) + g(x) > -\infty$ is attained at x^*. Consider the algorithm, for $x_0 \in \mathbb{R}^p$ and

$$x_{k+1} = \mathrm{prox}_{g/L}\left(x_k - \frac{1}{L}\nabla f(x_k)\right). \tag{7}$$

Then x_k converges to a global minimum and we have for any $k \in \mathbb{N}$, $k > 0$,

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

If in addition $f + g$ is μ-strongly convex, we have in addition

$$\|x_{k+1} - x^*\|_2^2 \leq \frac{L}{L+\mu}\|x_k - x^*\|_2^2.$$

Proximal gradient algorithm

- More efficient way to handle nonsmooth convex functions.
- Easier to implment.
- Faster in practice (similar as gradient descent).
- Require to compute prox operators (not always possible).

Plan

1. Gradient descent algorithm

2. Nonsmooth analsysis

3. Subgradient descent

4. Composite optimization

5. Lower bounds and acceleration

O(1/k) convergence rate?

Can we do better? What is the limit?

A lower bound

Definition

A first order method to minimize a smooth convex function f when initiated at $x_1 = 0$, produces a sequence of points $(x_i)_{i \in \mathbb{N}}$ such that for any $k \in \mathbb{N}$,

$$x_{k+1} \in \mathrm{span}\left(\nabla f(x_0), \ldots, \nabla f(x_k)\right).$$

Theorem

Let $k \leq (d-1)/2$, $L > 0$. There exists a convex function f with L-Lipschitz gradient over $\mathbb{R}^d$, such that for any first order method satisfying definition (11),

$$\min_{1 \leq s \leq k} f(x_s) - f(x^*) \geq \frac{3L}{32} \frac{\|x_0 - x^*\|^2}{(k+1)^2}.$$

Lower bound: proof (from Bubeck's book)

For $h : \mathbb{R}^d \to \mathbb{R}$ we denote $h^* = \inf_{x \in \mathbb{R}^d} h(x)$.

For $k \leq d$ let $A_k \in \mathbb{R}^{d \times d}$ be the symmetric and tridiagonal matrix defined by

$$(A_k)_{i,j} = \begin{cases} 2, & i = j, i \leq k \\ -1, & j \in \{i-1, i+1\}, i \leq k, j \neq k+1 \\ 0, & \text{otherwise.} \end{cases}$$

We verify that $0 \preceq A_k \preceq 4I$ since for any $x \in \mathbb{R}^d$,

$$\begin{aligned} x^T A_k x &= 2\sum_{i=1}^{k} x(i)^2 - 2\sum_{i=1}^{k-1} x(i)x(i+1) = x(1)^2 + x(k)^2 + \sum_{i=1}^{k-1}(x(i) - x(i+1))^2 \\ &\leq 4\sum_{i=1}^{k} x(i)^2 \geq 0 \end{aligned}$$

Lower bound: proof (from Bubeck's book)

$$A_k = \begin{pmatrix} \left.\begin{matrix} 2 & -1 & 0 & & & \\ -1 & 2 & -1 & & 0 & \\ & \dots & & & \dots & \\ & 0 & & -1 & 2 & -1 \\ & & & 0 & -1 & 2 \end{matrix}\right\} & k \text{ lines} \\ & \\ 0_{d-k,k} & 0_{d-k,d-k} \end{pmatrix}$$

We consider now the following convex function:

$$f(x) = \frac{L}{8} x^\top A_{2k+1} x - \frac{L}{4} x^\top e_1.$$

For any $s = 1, \dots, k$, x_s must lie in the linear span of $e_1, \dots, e_{s-1}$ (assumption). In particular for $s \le k$, $x_s(i) = 0$ for $i = s, \dots, d$, which implies $x_s^T A_{2k+1} x_s = x_s^T A_k x_s$. Set

$$f_k(x) = \frac{L}{8} x^\top A_k x - \frac{L}{4} x^\top e_1,$$

We proved that, for all $s \le k$

$$f(x_s) - f^* = f_k(x_s) - f^*_{2k+1} \ge f^*_k - f^*_{2k+1}.$$

Thus it remains to compute the minimizer x^*_k of f_k, its norm, and the corresponding function value f^*_k.

The point x_k^* is the unique solution in the span of $e_1, \ldots, e_k$ of $A_k x = e_1$. One can verify (Exercise) that it is defined by $x_k^*(i) = 1 - \frac{i}{k+1}$ for $i = 1, \ldots, k$. Thus we have:

$$f_k^* = \frac{L}{8}(x_k^*)^\top A_k x_k^* - \frac{L}{4}(x_k^*)^\top e_1 = -\frac{L}{4}(x_k^*)^\top e_1 = -\frac{L}{4}\left(1 - \frac{1}{k+1}\right).$$

Furthermore note that

$$\|x_k^*\|^2 = \sum_{i=1}^{k}\left(1 - \frac{i}{k+1}\right)^2 = \sum_{i=1}^{k}\left(\frac{i}{k+1}\right)^2 \leq \frac{k+1}{3}.$$

Thus one obtains:

$$f_k^* - f_{2k+1}^* = \frac{L}{4}\left(\frac{1}{k+1} - \frac{1}{2k+2}\right) \geq \frac{3L}{16}\frac{\|x_{2k+1}^*\|^2}{(k+1)^2},$$

Optimality of the lower bound?

Gradient descent achieves $1/k$ and the lower bound is $1/k^2$. Which one is tight?

Nesterov's accelerated method

Theorem

Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ be convex continuously differentiable with L-Lipschitz gradient $\inf_{x \in \mathbb{R}^p} f(x) > -\infty$. *Consider the algorithm, for $x_{-1} \in \mathbb{R}^p$, set $y_0 = x_{-1}$, $t_1 = 1$ and for $k \in \mathbb{N}$,*

$$\begin{aligned} x_k &= y_k - \frac{1}{L} \nabla f(y_k) \\ t_{k+1} &= \frac{1 + \sqrt{1 + 4t_k^2}}{2} \\ y_{k+1} &= x_k + \left(\frac{t_k - 1}{t_{k+1}} \right) (x_k - x_{k-1}). \end{aligned} \tag{8}$$

Then for any $k \in \mathbb{N}$

$$f(x_k) - f^* \leq \frac{4L\|x_0 - x^*\|_2^2}{(k+2)^2}.$$

Nesterov's acceleration: proof

Set for any $k \in \mathbb{N}$,

$$p_k := (t_k - 1)(x_{k-1} - x_k) \qquad \text{so that} \qquad y_{k+1} = x_k - \frac{p_k}{t_{k+1}}$$

Momentum term: for any $k \geq 1$

$$t_k \geq \frac{1 + \sqrt{4t_{k-1}^2 + 1}}{2} \geq t_{k-1} + \frac{1}{2} \geq t_0 + \frac{k}{2} = 1 + \frac{k}{2}. \tag{9}$$

$$(t_{k+1}^2 - t_{k+1}) = t_k^2. \tag{10}$$

Main argument: the sequence $\{z_k\}_{k\in\mathbb{N}}$,

$$z_k := \frac{2t_k^2}{L}(f(x_k) - f^*) + \|p_k - x_k + x^*\|^2, \tag{11}$$

is non-increasing and $z_0 \leq 2||x_0 - x^*||^2$. The result can be deduced by combining (9) and (11).

Nesterov's acceleration: proof

We have a series of three inequalities.

$$p_{k+1} - x_{k+1} = p_k - x_k + \frac{t_{k+1}}{L} \nabla f(y_{k+1})$$

This implies

$$\begin{aligned}\|p_{k+1} - x_{k+1} + x^*\|_2^2 &= \|p_k - x_k + x^*\|_2^2 + 2\frac{(t_{k+1} - 1)}{L} \langle p_k, \nabla f(y_{k+1}) \rangle \\ &\quad + 2\frac{t_{k+1}}{L} \langle x^* - y_{k+1}, \nabla f(y_{k+1}) \rangle + \frac{t_{k+1}^2}{L^2} \|\nabla f(y_{k+1})\|_2^2\end{aligned}$$

From the Lipschitz gradient assumption, we obtain

$$\begin{aligned}f(x_{k+1}) - f^* &\leq f(y_{k+1}) - f^* - \frac{1}{2L} \|\nabla f(y_{k+1})\|_2^2 \\ &\leq \langle \nabla f(y_{k+1}), y_{k+1} - x^* \rangle - \frac{1}{2L} \|\nabla f(y_{k+1})\|_2^2 \\ \frac{1}{2L} \|\nabla f(y_{k+1})\|_2^2 \leq f(y_{k+1}) - f(x_{k+1}) &\leq f(x_k) - f(x_{k+1}) - \frac{1}{t_{k+1}} \langle p_k, \nabla f(y_{k+1}) \rangle\end{aligned}$$

Using the last three identities, we obtain

$$
\begin{aligned}
&\|p_{k+1} - x_{k+1} + x^*\|_2^2 - \|p_k - x_k + x^*\|_2^2 \\
&= 2\frac{(t_{k+1}-1)}{L}\langle p_k, \nabla f(y_{k+1})\rangle + 2\frac{t_{k+1}}{L}\langle x^* - y_{k+1}, \nabla f(y_{k+1})\rangle + \frac{t_{k+1}^2}{L^2}\|\nabla f(y_{k+1})\|_2^2 \\
&\leq 2t_{k+1}\frac{(t_{k+1}-1)}{L}\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L}\|\nabla f(y_{k+1})\|_2^2\right) \\
&\quad + 2\frac{t_{k+1}}{L}\left(f^* - f(x_{k+1}) - \frac{1}{2L}\|\nabla f(y_{k+1})\|_2^2\right) + \frac{t_{k+1}^2}{L^2}\|\nabla f(y_{k+1})\|_2^2 \\
&= 2t_{k+1}\frac{(t_{k+1}-1)}{L}(f(x_k) - f^* + f^* - f(x_{k+1})) + 2\frac{t_{k+1}}{L}(f^* - f(x_{k+1})) \\
&= 2\frac{t_k^2}{L}(f(x_k) - f^*) - 2\frac{t_{k+1}^2}{L}(f(x_{k+1} - f^*))
\end{aligned}
$$

This proves that the sequence $(z_k)_{k\in\mathbb{N}}$ is non increasing. It remains to compute z_0,

$$z_0 = \frac{2}{L}(f(x_0) - f^*) + \|x^* - x_0\|^2 \leq 2\|x_0 - x^*\|_2^2.$$

Putting things together

$$f(x_k) - f^* \leq \frac{Lz_0}{2t_k^2} \leq \frac{4L\|x_0 - x^*\|_2^2}{(k+2)^2}.$$

Conclusion for nonsmooth optimization

- When available prefer proximal method.
- Acceleration works well in practive.
- Extension of Nesterov's algorithm to the proximal decomposition setting (Beck and Teboulle, FISTA).