Concise Lecture Notes on Optimization Methods for Machine Learning and Data Science

These lecture notes are publicly available but their use for teaching or even research purposes requires citing:

L. N. Vicente, S. Gratton, and R. Garmanjani, Concise Lecture Notes on Optimization Methods for Machine Learning and Data Science, ISE Department, Lehigh University, January 2019.

If appropriate, the corresponding source references given at the end of these notes should be cited instead.

These lecture notes are displayed in the form of slides for teaching convenience.

The problems typically addressed in ML/DS are of the form

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is smooth (∇f is at least Lipschitz continuous) and $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex (and proper and closed) typically non-smooth.

Examples:

1. Unconstrained optimization ($g = 0$).
2. Structured regularization, where g is a regularizer like the ℓ_1 one ($g(x) = \lambda \|x\|_1$, $\lambda > 0$).
3. Convex constrained smooth optimization, such as

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in C, \end{array}$$

with $C \neq \emptyset$ closed and convex, can be formulated with $g = \delta_C$ (indicator function of C)

$$\delta_C(x) \; = \; \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C \end{cases}$$

However only simple constraints are handled well in such a way.

A data set for analysis involving optimization is typically of the form

$$D \; = \; \{(a_j, y_j), j = 1, \ldots, N\}$$

where the a_j's vectors are features or attributes

the y_j's vectors are labels or observation or responses.

The analysis consists of finding a prediction function ϕ such that

$$\phi(a_j) \simeq y_j, \quad j = 1, \ldots, N$$

in some optimal sense.

Now

1. The process of finding ϕ is called learning or training.

2. When the y_j's are reals, one has a regression problem.

3. When the y_j's lie in a finite set $\{1, \ldots, M\}$, one has a classification problem.

   $M = 2$ leads to binary classification.

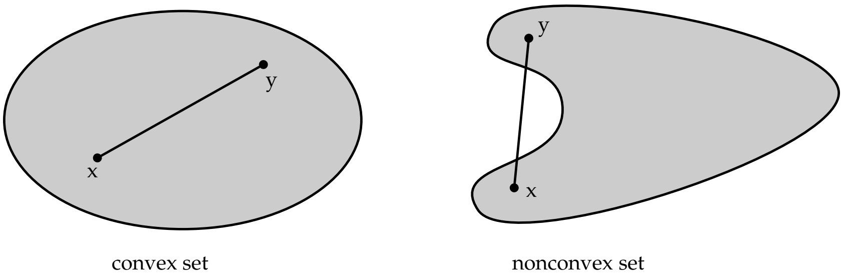
4. The labels may be null. In that case, one may want to group the a_j's in clusters (clusterization) or identify a low-dimensional subspace (or a collection of) where the a_j's lie (subspace identification).

   The labels may have to be learned while learning ϕ.

Convexity is a key concept in Optimization

A set C is convex if

$$\alpha x + (1-\alpha)y \in C, \quad \forall x, y \in C, \alpha \in [0,1]$$



$\alpha x + (1-\alpha)y, \alpha \in [0,1]$ is a convex combination of x and y.

$\sum_{i=1}^{n} \alpha_i x_i$ with $\sum_{i=1}^{n} \alpha_i = 1$, $\alpha_i \geq 0, \forall i$ is a convex combination of $x_1, \ldots, x_n$.

Examples

- $\mathbb{R}^n$
- $\emptyset$ (by convention)
- a subspace
- a polyhedral set $\{x \in \mathbb{R}^n : A_1 x = b_1, A_2 x \geq b_2\}$
    - a hyperplane $\{x \in \mathbb{R}^n : a^\top x = b\}$
    - a halfspace $\{x \in \mathbb{R}^n : a^\top x \geq b\}$
    - a system of linear equations $\{x \in \mathbb{R}^n : Ax = b\}$
    - a polytope (bounded polyhedral set)
- a convex cone: K is a cone if $\alpha x \in K, \forall \alpha > 0, x \in K$

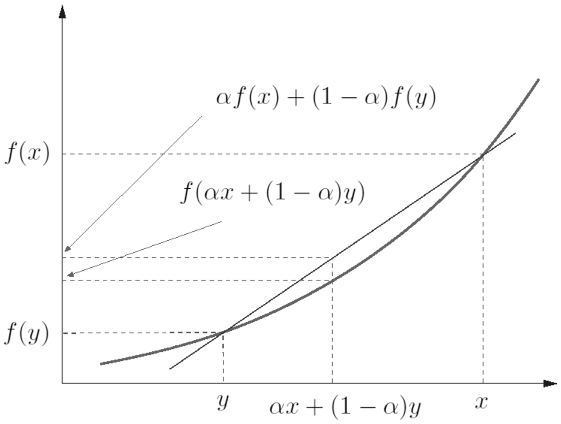
Operations preserving convexity

- Intersection $C_1 \cap C_2$
- Set sum $C_1 + C_2 = \{x_1 + x_2 : x_1 \in C_1, x_2 \in C_2\}$
- Affine transformation $f(C) = \{Ax + b : x \in C\}$ with $f(x) = Ax + b$. Particular cases:
  - scaling
  - translation
  - projection $\{x_1 : \exists x_2 : (x_1, x_2) \in C\}$
- Cartisian product $C_1 \times C_2 = \{(x_1, x_2) : x_1 \in C_1, x_2 \in C_2\}$

The proofs are left as EXERCISES.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if

$$f(\alpha x + (1-\alpha)y) \ \leq \ \alpha f(x) + (1-\alpha) f(y), \forall \alpha \in [0,1], x, y \in \mathbb{R}^n$$



Examples

- All norms: in particular p–norms, $\|x\|_p, 1 \leq p \leq \infty$.
- Affine functions: $f(x) = a^\top x + b$.

A convex function f could be of extended type $f : \mathbb{R}^n \to [-\infty, \infty]$.

An example is the indicator of a (convex) set C (seen before):

$$\delta_C(x) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C \end{cases}$$

Another example is $(n = 1)$

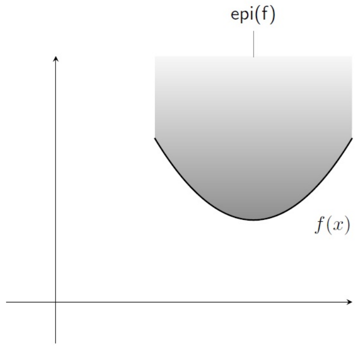
$$f(x) = -\log(x) \text{ if } x > 0, \quad f(x) = \infty \text{ if } x \leq 0.$$

We thus restrict our attention to proper convex functions where $-\infty$ is never attained and $\exists x \in \mathbb{R}^n : f(x) < \infty$ ($\Longrightarrow \mathrm{dom}(f) \neq \emptyset$).

Also, the definition of a convex function can be restricted to a convex set or to the domain of the function (assumed convex). Hopefully each example will reveal whether we are assuming convexity over a set, the domain, or the whole space.

In many examples it is difficult to apply directly the definition to verify convexity. Hence we use necessary and sufficient conditions:

1. (GEOMETRY) f is convex iff its epigraph $\mathrm{epi}(f) = \{(x,y) : y \geq f(x)\}$ is convex in $\mathbb{R}^{n+1}$



2. (REDUCTION TO THE SCALAR CASE) f is convex iff $f(x+\alpha v)$ is convex $\forall x, v$.

With this characterization it is easy to verify that

$$f(X) = -\log(\det(X))(= -\sum_{i=1}^{n} \log(\lambda_i(X)))$$

is convex in the space of matrices $n \times n$ and of domain $\{X \in \mathbb{R}^{n \times n} : X = X^\top, \quad \underbrace{\lambda_i(X)}_{i\text{–th eigenvalue of } X} \quad > 0, i = 1, \dots, n\}$

3. (CONTINUOUS DIFFERENTIABILITY) Let f be continuous differentiable in $\mathbb{R}^n$.

   f is convex iff $(\nabla f(y) - \nabla f(x))^\top (y - x) \geq 0 \; \forall x, y \in \mathbb{R}^n$

4. (TWICE CONT. DIFF.) Let f be twice cont. differentiable in $\mathbb{R}^n$.
   f is convex iff $d^\top \nabla^2 f(x) d \geq 0, \forall x, d$.

EXERCISE: Adapt 1–4 to the case where f is defined over a convex set C. In 3–4, you have to consider the convex cone of feasible directions: $A_C(x) = \{d \in \mathbb{R}^n : x + \alpha d \in C$ for some $\alpha > 0\}$.

It is easy to see that $f(x) = b^\top x + \frac{1}{2} x^\top A x$, with A symmetric and $\underbrace{\text{positive semi-definite}}_{\text{PSD}}$ (eigenvalues ≥ 0) is convex using 4: $\nabla^2 f(x) = A$ and $d^\top A d \geq 0, \forall d$.

Operations preserving convexity:

- Positive weighted sum $\sum_{i=1}^{p} \alpha_i f_i$, $\alpha_i > 0$, and f_i convex $\forall i$.
- Composition by affine transformation: $f(Ax + b)$.
- Pointwise maximum $\max_{1 \leq i \leq p} f_i(x)$, f_i convex $\forall i$.
- Composition by nondecreasing convex function

$$g(f), \ f, g \text{ convex and } g \text{ nondecreasing}$$

- Minimum over a closed convex set $g(x) = \inf_{y \in C} f(x, y)$

The proofs are left as EXERCISES.

One can now list more examples of convex functions relevant for this course and other data science contexts.

- $\|Ax - b\|_p$ and $\|Ax - b\|_2^2$
- $\sum_{i=1}^{p} e^{g_i(x)}$ with g_i convex
- $-\log(\det(X))$ for X symmetric PD (seen before)
- $-\sum_{i=1}^{p} \log(b_i - a_i^\top x)$
- distance to a set C (convex and closed)

$$d_C(x) \ = \ \min_{y \in C} \|x - y\| \quad \|\cdot\| = \|\cdot\|_2 \text{ by default}$$

- largest singular value of a matrix
- largest eigenvalue of a symmetric matrix

The proofs are left as EXERCISES.

Why is convexity relevant in Optimization?

Let us consider an optimization problem of the form

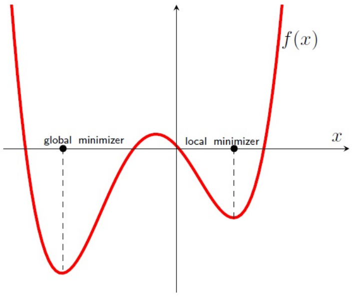
$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in \Omega \end{aligned}$$

A point x_* is a local (strict) minimizer if $\exists N$ neighborhood of x_* such that

$$f(x_*) \underset{(<)}{\leq} f(x), \quad \forall x \in (N \cap \Omega) \setminus \{x_*\}$$

x_* is said a global (strict) minimizer if

$$f(x_*) \underset{(<)}{\leq} f(x), \quad \forall x \in \Omega \setminus \{x_*\}$$



Then we have

Theorem

If f is convex over C, then every local minimum is global.

Proof.

If x a local minimizer is not global, $\exists z : f(z) < f(x)$. Then for $\alpha \in (0, 1)$:

$$\begin{aligned} f(\alpha z + (1-\alpha)x) &\leq \alpha f(z) + (1-\alpha) f(x) \\ &< f(x) \end{aligned}$$

which when $\alpha \to 0$ contradicts the fact that x is a local minimizer. □

Moreover, if f is strictly convex on C ("$<$" in the definition) and $\exists$ a local minimizer then $\exists$ a unique global minimizer. Why?

The set of minimizers of a convex function is convex. Why?

Convexity or strict convexity does not guarantee the existence of minimizers (take e^x in $\mathbb{R}$).

In the general, possible nonconvex case existence of minimizers is guaranteed by the Weierstrass Theorem: A continuous function has a minimizer (and a maximizer) in a compact set Ω (in $\mathbb{R}^n$ closed and bounded).

One can trade boundedness of Ω by uniform convexity of f in Ω convex.

A function f is uniformly convex (with constant $\mu_f = 0$ called the modulus) if

$$f(\alpha x + (1-\alpha)y) \ \leq \ \alpha f(x) + (1-\alpha) f(y) - \frac{\mu_f}{2}\alpha(1-\alpha)\|x-y\|^2$$

The smooth characterizations of uniform convexity are:

$$\begin{aligned} (\nabla f(y) - \nabla f(x))^\top (y-x) \ &\geq \ \mu_f \|x-y\|^2, \quad \forall x, y \\ d^\top \nabla^2 f(x) d \ &\geq \ \mu_f \|d\|^2, \quad \forall x, d \end{aligned}$$

Hence a quadratic $q(x) = b^\top x + \frac{1}{2} x^\top A x$ with A symmetric and PD is

μ_q–uniformly convex with $\mu_q = \lambda_{\min}(A)$.

Uniform convexity can be restricted to a convex set C (work out the details!).

A number of convex functions in this course are nonsmooth, and it is time now to define tools to deal with nondifferentiability.

Let f a be convex function, possibly of value extended to $[-\infty, +\infty]$.

The vector v is a subgradient of f at x if

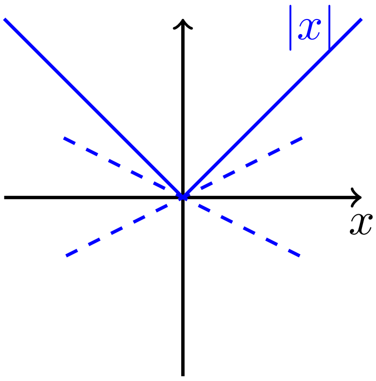
$$f(x+d) \; \geq \; f(x) + v^\top d, \quad \forall d \in \mathbb{R}^n.$$

The set of all subgradients is called subdifferential

$$\partial f(x) \; = \; \{v \in \mathbb{R}^n : v \text{ is subgradient of } f \text{ at } x\}$$

Let us see two examples:

$f(x) = |x|$. In this case $\partial f(0) = [-1, 1]$



$\delta_C(x) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C \end{cases}$. In this case, for any $x \in C$, $v \in \partial \delta_C(x)$ if

$$\delta_C(z) \;\geq\; \delta_C(x) + v^\top(z - x) \quad \forall z \in C$$

$$\Downarrow$$

$$v^\top(z - x) \;\leq\; 0 \quad \forall z \in C$$

$$\Updownarrow$$

$$v \in N_C(x)$$

where N_C is the cone normal to C at x.

What are the features of the subdifferential?

Theorem

If f is convex and proper then $\partial f(x)$ is closed and convex for all $x \in \operatorname{dom}(f)$.

Proof.

A simple consequence of $\partial f(x)$ being the intersection of half-spaces (which are closed and convex). □

Theorem

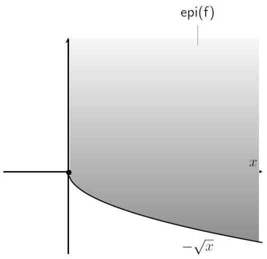
If f is convex and proper then $\underbrace{\partial f(x) \neq \emptyset}_{f \text{ is subdifferentiable at } x}$ and $\partial f(x)$ is bounded for all $x \in \operatorname{int}(\operatorname{dom}(f))$.

Proof.

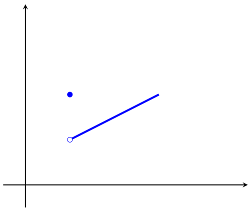
See Beck 2017. □

To better understand the last result, consider $f(x) = -\sqrt{x}$, $\mathrm{dom}(f) = \mathbb{R}_0^+$. Note that f is also closed in the sense that $\mathrm{epi}(f)$ is closed.

However, $\partial f(0)$ is the empty set!



Also, convex functions are not necessarily continuous at boundary points of their domains, as we see from



and this can even happen when f is closed (but for $n > 1$).

The subdifferential characterizes optimality for convex functions:

Theorem

x_* is a (global) minimum of f convex iff $0 \in \partial f(x_*)$

Proof.

If x_* is a minimizer,

$$f(x_* + d) \geq f(x_*) \geq f(x_*) + 0^\top d \quad \forall d,$$

showing that

$$0 \in \partial f(x).$$

If $0 \in \partial f(x_*)$,

$$f(x_* + d) \geq f(x_*) \quad \forall d.$$

□

Calculus rules for ∂f (for simplicity all convex functions are assumed real value with domain $\mathbb{R}^n$; proofs are left as EXERCISES).

Continuous differentiability $\partial f(x) = \{\nabla f(x)\}$

Positive weighted sum $\partial(\sum_{i=1}^{p} \alpha_i f_i)(x) = \sum_{i=1}^{p} \alpha_i \partial f_i(x)$

Composition by affine transformation $g = f(Ax + b)$

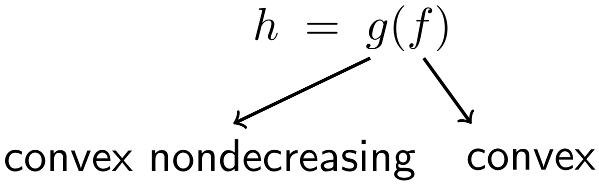
$$\partial g(x) \;=\; A^\top \partial f(Ax + b)$$

Pointwise maximum $g(x) = \max_{1\leq i\leq p} f_i(x)$

$$\partial g(x) = \underbrace{\mathrm{conv}}_{\text{convex hull}} \bigcup_{i\in I(x)} \partial f_i(x)$$

where $I(x) = \{i : f_i(x) = g(x)\}$. Hence, in a weak sense, any element of $\partial f_i(x)$, $i \in I(x)$, is in $\partial g(x)$.

Composition by nondecreasing convex function


$h \;=\; g(f)$

convex nondecreasing

convex

Let us assume that $g : \mathbb{R} \to \mathbb{R}$ is continuous differentiable. Then

$$\partial h(x) \;=\; g'(f(x))\partial f(x)$$

Distance to a closed convex set $d_C(x) = \min_{y \in C} \|x - y\|$

$$\partial d_C(x) \; = \; \left\{ \frac{x - P_c(x)}{\underbrace{d_C(x)}_{=\|x-P_C(x)\|}} \right\} \quad \text{for} \quad x \notin C$$

where $P_C(x)$ is the orthogonal projection of x onto C.

If $x \in C$, $\partial d_C(x) = \underbrace{N_C(x)}_{\text{normal cone}} \cap \underbrace{B(0;1)}_{\{v:\|v\|\leq 1\}}$.

In particular, one has $0 \in \partial d_C(x)$.

Examples

- $\partial f(x) = \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1$

$$\begin{aligned}\partial f(x) &= A^\top Ax + A^\top b + \lambda\partial_{\|\cdot\|}(x)\\ &= A^\top Ax + A^\top b + \lambda\left\{\sum_{x_i\neq 0}\mathrm{sign}(x_i)e_i + \sum_{x_i=0}[-e_i, e_i]\right\}\end{aligned}$$

  with $I = [e_1 \dots e_n]$.

- $f(x) \;=\; \sum_{i=1}^p |a_i^\top x - b_i|$

$$\partial f(x) \;=\; \sum_{i=1}^p \partial_{|\cdot|}(a_i^\top x - b_i)a_i$$

- $f(x) = \|x\|_2$

$$\partial f(x) = \begin{cases} x/\|x\|_2, & x \neq 0\\ B(0;1), & x = 0\end{cases}$$

NOTE: $\nabla f(x_*) = 0$ when f is nonconvex and continuous differentiable is still a necessary condition, but not longer sufficient.

We end this background chapter with some notions of rates of convergent sequences. We will be interested in knowing the speed of convergence of sequences such as

$f(x_k) - f(x_*)$ optimality gap (when x_* is a minimizer)

$$\left.\begin{array}{l} \|\nabla f(x_k)\| \\ d_{\partial f(x_k)}(0) \\ \end{array}\right\} \begin{array}{l} \text{stationary or criticality (smooth and} \\ \text{nonsmooth case, respectively)} \end{array}$$

$\|x_k - x_*\|$ absolute error in the iterates (when $x_k \to x_*$)

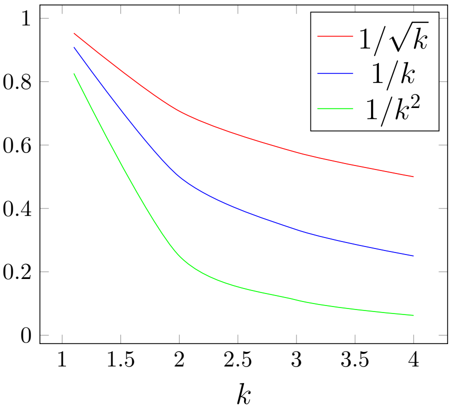
Let $\{\omega_k\} \subset \mathbb{R}^n$ be a sequence converging to ω_*. There are four major types of rates of convergence of interest to us

- SUBLINEAR $\lim_{k\to\infty} \frac{\|\omega_{k+1}-\omega_*\|}{\|\omega_k-\omega_*\|} = 1$

  When $n = 1$ and $\omega_* = 0$, at least three examples will be seen in this course:

$$\frac{1}{\sqrt{k}}, \quad \frac{1}{k}, \quad \frac{1}{k^2}$$

  Sublinear is a slow rate but $\frac{1}{k^2}$ is much faster than $\frac{1}{\sqrt{k}}$.



- LINEAR (also known as geometric or exponential convergence):
  $\exists r \in (0,1) \quad \|\omega_{k+1} - \omega_*\| \leq r\|\omega_k - \omega_*\| \quad \forall k$

  When $n = 1$ and $\omega_* = 0$, an example is $(\frac{1}{2})^k$.

  First-order methods (like the gradient or steepest descent method) exhibit sublinear or linear rates. Second-order methods achieve a sublinear rate (quasi-Newton) or a quadratic rate (Newton).

- SUPERLINEAR $\exists \{\eta_k\}_{\eta_k \to 0} \quad \|\omega_{k+1} - \omega_*\| \leq \eta_k\|\omega_k - \omega_*\| \quad \forall k$

  The example for $n = 1$ and $\omega_* = 0$ is $\frac{1}{k!}$

- QUADRATIC $\exists M > 0, \|\omega_{k+1} - \omega_*\| \leq M\|\omega_k - \omega_*\|^2 \quad \forall k$

  Take $10^{(1/2)^k}$ as the example when $n = 1$ and $\omega_* = 0$.

Quadratic $\Longrightarrow$ Superlinear $\Longrightarrow$ Linear $\Longrightarrow$ Sublinear Why?

Somehow we have presented a local version of these rates since (by assuming that $\omega_k \to \omega_*$) we supposed that ω_0 is sufficiently close to ω_*.

These rates are called global when no assumption is made about ω_0.

Also the version presented is what is known as the "q–rates".

See, from $\omega_{k+1} \leq \frac{1}{2}\omega_k$ $(\omega_* = 0, \omega_k > 0 \ \forall k)$ one has

$$\omega_k \ \leq \ \left(\frac{1}{2}\right)^k \omega_0.$$

However not all sequences satisfying this rate are q–linear.

A trivial example is

$$\omega_k = \begin{cases} (\frac{1}{2})^k & \text{when } k \text{ is even} \\ 0 & \text{when } k \text{ is odd} \end{cases}$$

Such $\{\omega_k\}$ converges r–linearly to 0.

In general a sequence $\{\omega_k\} \subset \mathbb{R}^n, \omega \to \omega_*$, has a r–linear rate if $\|\omega_k - \omega_*\|$ is bounded by a sequence in $\mathbb{R}$ that converges q–linearly to zero.

For the minimization of a smooth (cont. diff.) function f in $\mathbb{R}^n$, the steepest descent or gradient method is

$$x_{k+1} \;=\; x_k - \alpha_k \nabla f(x_k)$$

where $\alpha_k > 0$ is the step size.

The negative gradient $-\nabla f(x)$ is a descent direction

$$\Big\downarrow$$

$$d \neq 0 : f'(x; d) \;<\; 0$$

$$\Downarrow$$

$$\exists \bar{\epsilon} > 0 : f(x + \epsilon d) \;<\; f(x), \quad \forall \epsilon \in (0, \bar{\epsilon}]$$

In fact

$$f'(x; -\nabla f(x)) \;=\; \nabla f(x)^\top (-\nabla f(x)) \;=\; -\|\nabla f(x)\|^2 \;<\; 0$$

The exact line search strategy consists of choosing

$$\alpha_k \;=\; \mathrm{argmin}_{\alpha>0}\, f(x_k - \alpha\nabla f(x_k)).$$

Other strategies will be covered in the next chapter.

When f is not differentiable, $-\nabla f(x_k)$ may not exist. We thus assume that f is convex and $\mathrm{int}(\mathrm{dom}(f)) = \mathbb{R}^n$, and consider a generalization called the subgradient method

$$x_{k+1} \;=\; x_k - \alpha_k g_k, \quad g_k \in \partial f(x_k),$$

where the subgradient g_k is in the subdifferential $\partial f(x_k)$.

Importantly, $-g_k$ might not be a descent direction!

Example

$f(x_1, x_2) = |x_1| + 2|x_2|$

$$\partial f(1,0) \;=\; \{(1,x) : |x| \le 2\}$$

$$(1,2) \in \partial f(1,0)$$

$$-(1,2) \text{ is not descent : } f'((1,0); -(1,2)) \;=\; g'_+(0) \;=\; 3 > 0$$

with

$$g(\alpha) = f((1,0) - \alpha(1,2)) \;=\; |1-\alpha| + 4\alpha \;=\; \begin{cases} 1 + 3\alpha, & \alpha \in [0,1] \\ 5\alpha - 1, & \alpha \ge 1 \end{cases}$$

Notes:

1. f is subdifferentiable over C ($\partial f(x) \neq \emptyset, \forall x \in C$) as it is convex and $C \subseteq \mathrm{int}(\mathrm{dom}(f))$ (see Chapter 2).
2. The orthogonal projection over C is Lips. continuous with constant 1 (nonexpensive): $\|P_C(x) - P_C(y)\| \leq \|x - y\|, \forall x, y.$ Why?
3. If $g_k = 0$ for some k, then x_k is a minimizer and $x_i = x_k \forall i \geq k$.

A fundamental inequality for projected subgradient is

$$\boxed{\|x_{k+1} - x_*\|^2 \ \leq\ \|x_k - x_*\|^2 - 2\alpha_k(f(x_k) - f_*) + \alpha_k^2\|g_k\|^2,}$$

$\forall x_* \in X_*$ (the set of minimizers of f in C assumed $\neq \emptyset$, which is necessarily closed). Why?

f_* is the optimal value.

Proof.

$$\begin{aligned}
\|x_{k+1} - x_*\|^2 &= \|P_C(x_k - \alpha_k g_k) - P_C(x_*)\|^2 \\
&\leq \|x_k - \alpha_k g_k - x_*\|^2 \\
&= \|x_k - x_*\|^2 - 2\alpha_k g_k^\top (x_k - x_*) + \alpha_k^2 \|g_k\|^2 \\
&\overset{\text{subgradient inequality}}{\leq} \|x_k - x_*\|^2 - 2\alpha_k (f(x_k) - f_*) + \alpha_k^2 \|g_k\|^2
\end{aligned}$$

□

A natural choice for α_k is the minimizer of the RHS of the fundamental inequality for $\alpha \geq 0$

$$\alpha_k \ = \ \frac{f(x_k) - f_*}{\|g_k\|^2}$$

This results in the Polyak's stepsize rule

$$\alpha_k = \begin{cases} \frac{f(x_k)-f_*}{\|g_k\|^2} & \text{if } g_k \neq 0 \\ 1 & \text{otherwise} \end{cases}$$

One will assume that

$$\boxed{\|g\| \;\leq\; L_{\partial f} \quad \forall g \in \partial f(x)\; \forall x \in C}$$

which actually implies that

$$|f(x) - f(y)| \;\leq\; L_{\partial f}\|x - y\| \quad \forall x, y \in C$$

Why?

Rate of convergence of Polyak's stepsize rule:

$$\boxed{f^k_{\text{best}} - f_* \;\leq\; \frac{L_{\partial f} d_{X_*}(x_0)}{\sqrt{k+1}}, \quad \forall k \geq 0}$$

where $f^k_{\text{best}} = \min_{0 \leq i \leq k} f(x_i)$ (and $d_{X_*}(x_0) < \infty$ since X_* is closed)

Proof.

Plugging the stepsize rule in the fundamental inequality:

$$\begin{aligned}
\|x_{k+1} - x_*\|^2 &\leq \|x_k - x_*\|^2 - \frac{(f(x_k) - f_*)^2}{\|g_k\|^2} \\
&\leq \|x_k - x_*\|^2 - \frac{(f(x_k) - f_*)^2}{L_{\partial f}^2}
\end{aligned}$$

Thus

$$\begin{aligned}
\frac{1}{L_{\partial f}^2} \sum_{i=0}^{k} (f(x_i) - f_*)^2 &\leq \|x_0 - x_*\| - \|x_{k+1} - x_*\|^2 \\
&\leq \|x_0 - x_*\|^2 \\
&\leq d_{X_*}^2(x_0)
\end{aligned}$$

and

$$(k+1)(f_{\text{best}}^k - f_*)^2 \leq L_{\partial f}^2 d_{X_*}^2(x_0).$$

Consequences of the proof:

- $\|x_{k+1} - x_*\| \leq \|x_k - x_*\|, \quad \forall k, \ \forall x_* \in X_*$
- $\lim_{k\to\infty} f(x_k) = f_*$

Property 1) is Fejér monotonicity of $\{x_k\}$ w.r.t X_* which actually implies that $\{x_k\}$ does converge to a point in X_*. EXERCISE (see Beck 2017)

The worst case complexity (WCC) is $\mathcal{O}(\epsilon^{-2})$ in the sense that $\mathcal{O}(\epsilon^{-2})$ iterations are required to obtain a x_k such that $f^k_{\text{best}} - f_* \leq \epsilon$. Why?

The limit $f(x_k) \to f_*$ holds for choices of α_k such that $(\sum_{i=0}^k \alpha_i^2)/(\sum_{i=0}^k \alpha_i) \xrightarrow[k\to\infty]{} 0$. EXERCISE (see Beck 2017). An example is $\alpha_k = \frac{1}{\sqrt{k+1}}$.

Variations of this also achieve the $1/\sqrt{k}$ rate (see Beck 2017).

EXERCISE

Apply the subgradient method to $f(x_1, x_2) = |x_1 + 2x_2| + |3x_1 + 4x_2|$.

Solution of linear feasibility problems

EXERCISE: State the alternating projection method when

$$S_1 = \{x \in \mathbb{R}^n : Ax = b\} \quad S_2 = \{x \in \mathbb{R}^n : x \geq 0\}$$

and then, alternatively, the greedy one when

$$\begin{aligned} S_i &= \{x \in \mathbb{R}^n : a_i^\top x = b_i, i = 1, \ldots, m\}, (a_i \text{ is the } i\text{-th row of } A) \\ S_{m+1} &= \{x \in \mathbb{R}^n : x \geq 0\}. \end{aligned}$$

Implement both for $A = \begin{pmatrix} 0 & 6 & -7 & 1 \\ -1 & 2 & 10 & -1 \end{pmatrix}$, $b = \begin{pmatrix} 0 \\ 10 \end{pmatrix}$, and plot $f(x_k)$, for $k = 1, \ldots, 20$, in both cases.

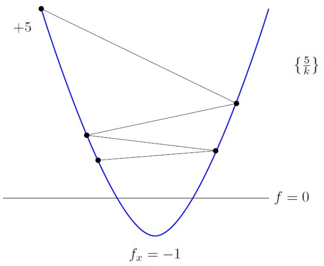
In which cases would you then consider the alternating one? ...

As we have seen before the gradient method, for continuous differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, is defined by

$$x_{k+1} \;=\; x_k - \alpha_k \nabla f(x_k)$$

where $\alpha_k > 0$ is the step size.

A choice of α_k that only ensures a simple decrease on f might not guarantee convergence to a stationary point



Thus, one has to ensure some form of sufficient decrease

$$f(x_k - \alpha_k \nabla f(x_k)) \;\leq\; f(x_k) - c\alpha_k \|\nabla f(x_k)\|_2^2$$

with $c \in (0, 1)$.

Note that when $x_{k+1} = x_k - \alpha_k p_k$ with p_k a descent direction ($f'(x_k; p_k) = -\nabla f(x_k)^\top p_k < 0$) sufficient decrease reads like

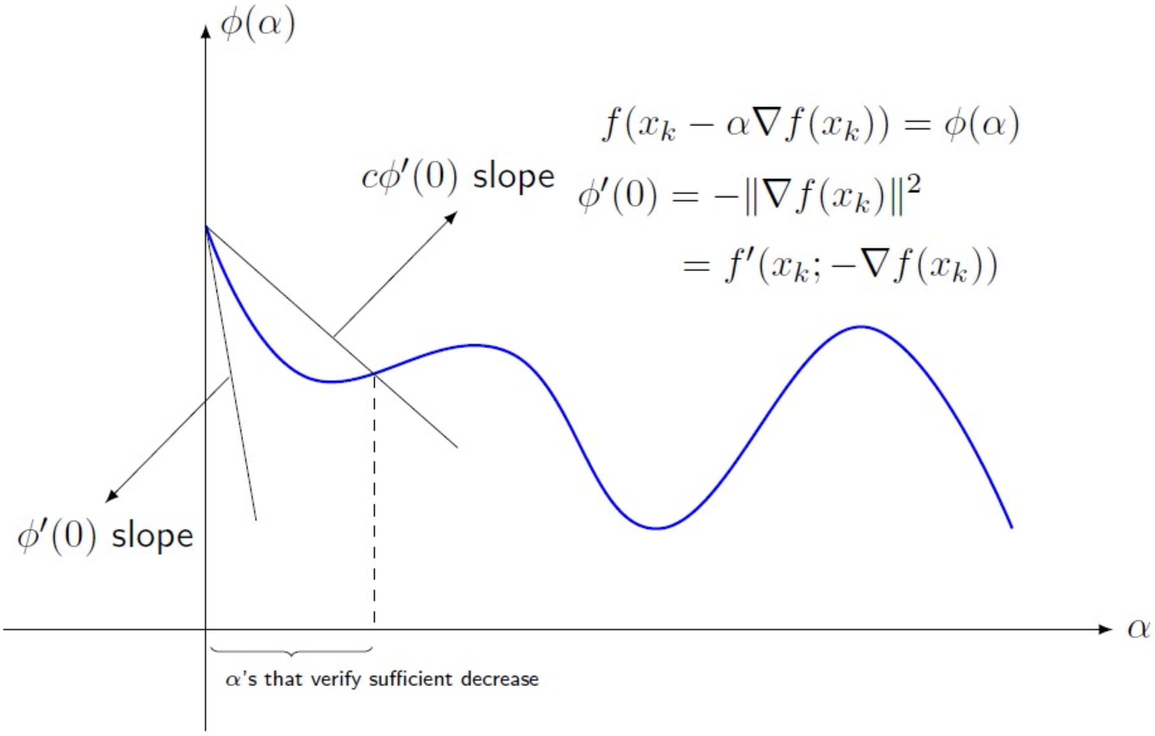
$$f(x_k - \alpha_k p_k) \;\leq\; f(x_k) - c\alpha_k \nabla f(x_k)^\top p_k$$

Such a sufficient decrease condition is typically imposed in Newton or quasi–Newton type methods.

Sufficient decrease guaranteed by a backtracking procedure:

```
Choose $c, \beta \in (0, 1)$ and $s > 0$. Set $\bar{\alpha} = s$.
while $f(x_k - \bar{\alpha}\nabla f(x_k)) \;>\; f(x_k) - c\bar{\alpha}\|\nabla f(x_k)\|^2$ do
    $\alpha := \beta \times \bar{\alpha}$
end while
$\alpha_k := \bar{\alpha}$
```

As long as f is bounded below, and thus bounded below in $\{x_k - \alpha \nabla f(x_k) : \alpha \geq 0\}$, this procedure may end in a finite number of steps, recalling, of course, that $-\nabla f(x_k)$ is a descent direction:


$\phi(\alpha)$
$f(x_k - \alpha \nabla f(x_k)) = \phi(\alpha)$
$c\phi'(0)$ slope
$\phi'(0) = -\|\nabla f(x_k)\|^2$
$= f'(x_k; -\nabla f(x_k))$
$\phi'(0)$ slope
α
α's that verify sufficient decrease

Let us assume now that ∇f is Lipschitz continuous with constant $L_{\nabla f} > 0$. Note that f may be nonconvex.

As we know from Chapter 2,

$$f(\overbrace{x_k - \alpha \nabla f(x_k)}^{y_k}) - f(x_k) \ \leq \ \nabla f(x_k)(\underbrace{-\alpha \nabla f(x_k)}_{y_k - x_k}) + \frac{L_{\nabla f}}{2} \| \underbrace{-\alpha \nabla f(x_k)}_{y_k - x_k} \|^2$$

giving rise to

$$f(x_k) - f(x_k - \alpha \nabla f(x_k)) \ \geq \ \alpha \left(1 - \frac{\alpha}{2} L_{\nabla f}\right) \|\nabla f(x_k)\|^2.$$

Besides, if $\beta\bar{\alpha}$ does not satisfy sufficient decrease

$$f(x_k) - f(x_k - (\beta\bar{\alpha})\nabla f(x_k)) \ < \ c(\beta\bar{\alpha})\|\nabla f(x_k)\|^2,$$

and this inequality together with the previous one with $\alpha = \beta\bar{\alpha}$ yield

$$\beta\bar{\alpha}\left(1 - \frac{\beta\bar{\alpha}}{2}L_{\nabla f}\right) \; < \; c(\beta\bar{\alpha})$$

thus

$$\bar{\alpha} \; > \; \frac{2(1-c)}{\beta L_{\nabla f}}$$

Hence

$$f(x_k) - f(\underbrace{x_{k+1}}_{x_k - \alpha_k \nabla f(x_k)}) \; \geq \; M\|\nabla f(x_k)\|^2$$

with $M = c\min\left\{s, \frac{2(1-c)}{\beta L_{\nabla f}}\right\}$.

Such an inequality is the key to analyze complexity and convergence for the gradient method with step size satisfying sufficient decrease.

In fact, summing this inequality from 0 to $k-1$ and noting the telescoping sum

$$f(x_0) - f(x_k) \;\geq\; M \sum_{i=0}^{k-1} \|\nabla f(x_i)\|^2$$

and, assuming a lower bound f_{low} on f, we reach the rate of convergence for such method

$$\boxed{\min_{0 \leq i \leq k-1} \|\nabla f(x_i)\| \;\leq\; \sqrt{\frac{f(x_0) - f_{low}}{M}} \, \frac{1}{\sqrt{k}}, \quad \forall k \geq 0}$$

as in the subgradient method for convex functions.

As a consequence of this proof the series $\sum_{i=0}^{\infty} \|\nabla f(x_i)\|^2$ is summable and therefore the gradient goes to zero

$$\lim_{k \to \infty} \nabla f(x_k) \;=\; 0$$

Moreover, the WCC is $\mathcal{O}(\epsilon^{-2})$ in the sense that $\mathcal{O}(\epsilon^{-2})$ iterations are required to obtain a x_* such that $\|\nabla f(x_*)\| \leq \epsilon$.

The sublinear rate $1/\sqrt{k}$ of the gradient method (with sufficient decrease) is, of course, slow and in the nonconvex case other line search methods based on Newton or quasi-Newton directions are much faster (quadratic or superlinear rates respectively) but require second-order information.

First-order methods (such as the gradient method) find room for application in problems where second-order information is prohibited such as those handling a large amount of data per iteration.

In those situations the function f is typically convex, an assumption made for the rest of this chapter.

The convex case will be treated in the more general scenario (see Chapter 1) where the problem is

$$\min_{x \in \mathbb{R}^n} F(x) \equiv f(x) + g(x)$$

still covering smooth unconstrained optimization ($g = 0$), but then addressing structured regularization ($g(x) = \lambda \|x\|_1$ for instance) and simple convex constraints ($g(x) = \delta_C$, where $C \neq \emptyset$ is closed and convex).

In addition, to later deal efficiently with the inclusion of g, we will cover gradient methods in their proximal variant.

And because of the features of most optimization data problems requiring regularization, proximal gradient methods will be analyzed only when f is convex or strongly convex.

It is simple to see that the gradient method can be expressed as

$$x_{k+1} \;=\; \operatorname*{argmin}_{x\in\mathbb{R}^n} \left\{ f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{1}{2\alpha_k}\|x - x_k\|^2 \right\}$$

in other words, as the minimizer of the sum of the linearization of f around x_k with a quadratic proximal term.

So, when dealing with the minimization of $F = f + g$, it is natural to consider

$$x_{k+1} \;=\; \operatorname*{argmin}_{x\in\mathbb{R}^n} \left\{ f(x_k) + \nabla f(x_k)^\top (x - x_k) + g(x) + \frac{1}{2\alpha_k}\|x - x_k\|^2 \right\}$$

Note that

$$\begin{aligned} &\frac{1}{2}\|x - (x_k - \alpha_k \nabla f(x_k))\|^2 \\ &= \; \frac{1}{2}\|x - x_k\|^2 + (x - x_k)^\top \left(\alpha_k \nabla f(x_k)\right) + (\text{constant in } x) \end{aligned}$$

Hence, the proximal subproblem (after multiplying its objective by α_k) can be written as

$$x_{k+1} \;=\; \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \alpha_k g(x) + \frac{1}{2} \|x - (x_k - \alpha_k \nabla f(x_k))\|^2 \right\}$$

We now introduce the proximal-operator (prox-operator):

$$\mathrm{prox}_{\alpha g}(x) \;=\; \mathrm{argmin}_{u \in \mathbb{R}^n} \left\{ \alpha g(u) + \frac{1}{2} \|u - x\|^2 \right\}$$

(for fixed α and x). The proximal gradient method can thus be written as

$$\boxed{x_{k+1} \;=\; \mathrm{prox}_{\alpha_k g}(x_k - \alpha_k \nabla f(x_k))}$$

where $\alpha_k > 0$ is the stepsize.

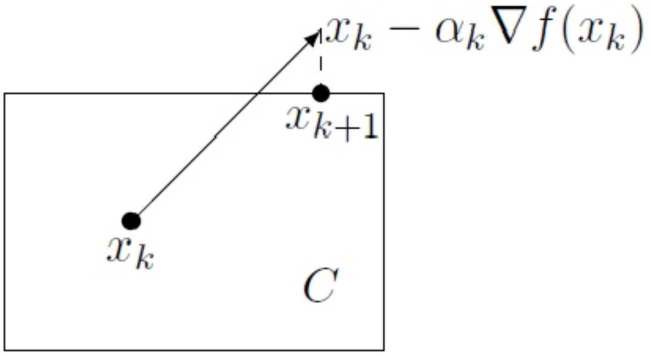
We now pause to cover examples and properties of the prox-operator. Let us start by three important examples:

1. $g(x) = 0, \forall x$. In this case, $\mathrm{prox}_{\alpha g}(x) = x$. This shows of course that the proximal gradient method reduces to the gradient one when there is no regularization.
2. $g(x) = \delta_C(x)$ with C closed and convex. Here

$$\begin{aligned} \mathrm{prox}_{\alpha g}(x) &= \operatorname*{argmin}_{u \in \mathbb{R}^n} \left\{ \alpha \delta_C(u) + \frac{1}{2}\|u - x\|^2 \right\} \\ &= \operatorname*{argmin}_{u \in C} \left\{ \frac{1}{2}\|u - x\|^2 \right\} \\ &= P_C(x) \end{aligned}$$

   which is simply the projection of x onto C.

   The proximal gradient method is then the projected gradient one.

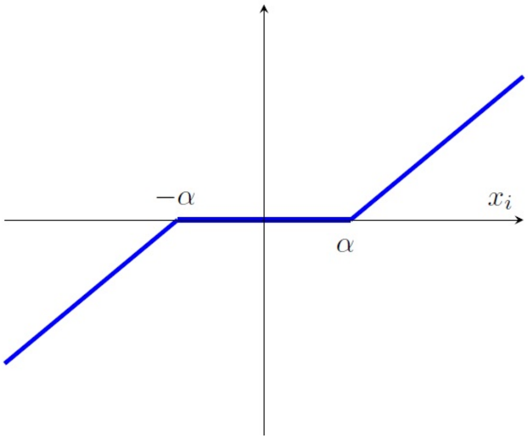


3. $h(x) = \alpha \|x\|_1$. One can then see that the minimization in the definition of the prox-operator separates in its n components, being the i-th one

$$\left(\mathrm{prox}_{\alpha\|\cdot\|}(x)\right)_i = \mathrm{argmin}_{u_i} \left\{ \alpha |u_i| + \frac{1}{2}(u_i - x_i)^2 \right\}$$

and thus

$$\left(\operatorname{prox}_{\alpha\|\cdot\|}(x)\right)_i = \begin{cases} x_i - \alpha & \text{if } x_i \geq \alpha \\ 0 & \text{if } x_i \in (-\alpha, \alpha) \\ x_i + \alpha & \text{if } x_i \leq \alpha \end{cases}$$

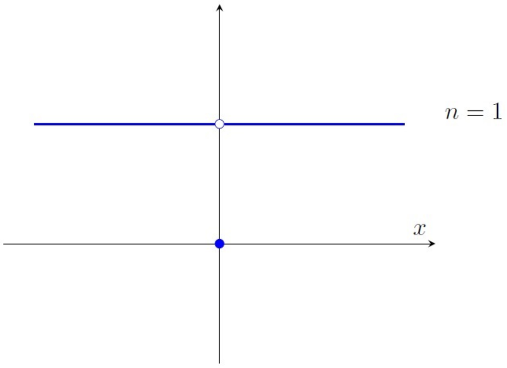


This is called the soft-thresholding operation.

A fourth example is hard-thresholding,

$$g(x) \;=\; \|x\|_0 \;=\; \left|\{i \in \{1,\ldots,n\} : x_i \neq 0\}\right|,$$

the number of nonzero components of x. Although g is not convex,



the prox-operator is well defined and separates into n components:

$$\left(\operatorname{prox}_{\alpha\|\cdot\|}(x)\right)_i \;=\; \begin{cases} x_i & \text{if } |x_i| \geq \sqrt{2\alpha}, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{Why?}$$

$i = 1, \ldots, n$.

We will analyze the proximal gradient method for $F = f + g$ when f is smooth (∇f Lipschitz continuous with constant $L_{\nabla f}$). Then

$$f(y) \;\leq\; f(x) + \nabla f(x)^\top (y - x) + \frac{L_{\nabla f}}{2} \|y - x\|^2, \quad \forall x, y. \qquad (B1)$$

As we said before, given the presence of g, the cases of interest are when f is convex or strongly convex. Then

$$f(y) \;\geq\; f(x) + \nabla f(x)^\top (y - x) + \frac{\mu_f}{2} \|y - x\|_2^2, \quad \forall x, y. \qquad (B2)$$

($\mu_f = 0$ convex; $\mu_f > 0$ strongly convex).

It will be very convenient to write the method as

$$\begin{aligned} x_{k+1} &= \mathrm{prox}_{\alpha_k g}(x_k - \alpha_k \nabla f(x_k)) \\ &= x_k - \alpha_k G_{\alpha_k}(x_k) \end{aligned}$$

$G_\alpha(x)$ is called the gradient mapping but it is not a gradient or a subgradient of $F = f + g$.

Moreover from $(*)$ and omitting the subscript k

$$(x - \alpha \nabla f(x)) - (x - \alpha G_\alpha(x)) \; \in \; \partial(\alpha g)\,(x - \alpha G_\alpha(x))$$

$$\Downarrow$$

$$\boxed{G_\alpha(x) - \nabla f(x) \; \in \; \partial g(x - \alpha G_\alpha(x)) \qquad (**)}$$

(Remark: From here one has that $G_\alpha(x) = 0 \iff x$ minimizes $F = f + g$.)

Our stepsize rule will be simply $\alpha = \frac{1}{L_{\nabla f}}$ but the derivation to come holds with $\alpha \in (0, \frac{1}{L_{\nabla f}}]$. From $(B1)$

$$f(y) \;\leq\; f(x) + \nabla f(x)^\top (y - x) + \frac{L_{\nabla f}}{2} \|y - x\|^2, \quad \forall x, y,$$

with $y = x - \alpha G_\alpha(x)$, one has then

$$f(x - \alpha G_\alpha(x)) \;\leq\; f(x) - \alpha \nabla f(x)^\top G_\alpha(x) + \frac{\alpha}{2} \|G_\alpha(x)\|^2 \quad (\mathsf{B1}_\alpha)$$

(Here we used $\alpha \leq 1/L_{\nabla f}$.)

We are now ready to prove a key inequality measuring the decrease along $G_\alpha(x)$. First we add $g(x - \alpha G_\alpha(x))$ to both sides of (B1$_\alpha$)

$$F(x - \alpha G_\alpha(x)) \;\leq\; f(x) - \alpha \nabla f(x)^\top G_\alpha(x) + \frac{\alpha}{2}\|G_\alpha(x)\|^2 + g(x - \alpha G_\alpha(x))$$

For any z, one has from $(B2)$

$$f(y) \;\geq\; f(x) + \nabla f(x)^\top (y - x) + \frac{\mu_f}{2}\|y - x\|^2$$

with y replaced by z

$$f(z) \;\geq\; f(x) + \nabla f(x)^\top (z - x) + \frac{\mu_f}{2}\|z - x\|^2$$

and using it above

$$
\begin{aligned}
F(x-\alpha G_\alpha(x)) \;\le\; & f(z)+g(z)-\nabla f(x)^\top(z-x)-\frac{\mu_f}{2}\|z-x\|^2 \\
& -\alpha\nabla f(x)^\top G_\alpha(x)+\frac{\alpha}{2}\|G_\alpha(x)\|^2+g(x-\alpha G_\alpha(x))-g(z) \\
=\; & F(z)-\nabla f(x)^\top(z-(x-\alpha G_\alpha(x))) \\
& +G_\alpha(x)^\top(z-(x-\alpha G_\alpha(x)))+G_\alpha(x)^\top(x-z) \\
& -\alpha\|G_\alpha(x)\|^2-\frac{\mu_f}{2}\|z-x\|^2+\frac{\alpha}{2}\|G_\alpha(x)\|^2 \\
& +g(x-\alpha G_\alpha(x))-g(z)
\end{aligned}
$$

From $(**)$

$$
g(x-\alpha G_\alpha(x))-g(z) \;\le\; (G_\alpha(x)-\nabla f(x))^\top(x-\alpha G_\alpha(x)-z).
$$

Hence the desired inequality

$$\boxed{\begin{array}{c} F(x-\alpha G_\alpha(x)) \ \leq \ F(z)+G_\alpha(x)^\top(x-z)-\frac{\alpha}{2}\|G_\alpha(x)\|^2-\frac{\mu_f}{2}\|z-x\|^2 \\ (***) \end{array}}$$

from which we can now extract rates of convergence for the convex ($\mu_f = 0$) and strongly convex ($\mu_f > 0$) cases.

First, we point out that setting $z = x$ shows us that the method is indeed descent

$$F(x-\alpha G_\alpha(x)) \ \leq \ F(x)-\frac{\alpha}{2}\|G_\alpha(x)\|^2.$$

Setting $z = x_*$ (a global minimizer of F)

$$\begin{aligned}
F(x - \alpha G_\alpha(x)) - F_* \;&\leq\; G_\alpha(x)^\top (x - x_*) - \frac{\alpha}{2}\|G_\alpha(x)\|^2 - \frac{\mu_f}{2}\|x - x_*\|^2 \\
&= \frac{1}{2\alpha}\left(\|x - x_*\|^2 - \|x - x_* - \alpha G_\alpha(x)\|^2\right) \\
&\quad - \frac{\mu_f}{2}\|x - x_*\|^2 \\
&= \frac{1}{2\alpha}\left((1 - \mu_f \alpha)\|x - x_*\|^2 - \|(x - \alpha G_\alpha(x)) - x_*\|^2\right)
\end{aligned}$$

and using again the indices k

$$\boxed{F(x_{k+1}) - F_* \;\leq\; \frac{1}{2\alpha_k}\left((1 - \mu_f \alpha_k)\|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2\right) \quad (B3)}$$

We can now derive a global rate.

In the convex case ($\mu_f = 0$), summing from 0 to k (with telescoping cancellation), and using $\alpha_k = 1/L_{\nabla f}$,

$$\begin{aligned}\sum_{i=1}^{k}(F(x_i) - F_*) &\leq \frac{L_{\nabla f}}{2}\sum_{i=1}^{k}\left(\|x_{i-1} - x_*\|^2 - \|x_i - x_*\|^2\right) \\ &= \frac{L_{\nabla f}}{2}\left(\|x_0 - x_*\|^2 - \|x_k - x_*\|^2\right) \\ &\leq \frac{L_{\nabla f}}{2}\|x_0 - x_*\|^2\end{aligned}$$

and because $F(x_i)$ is nondecreasing we arrive finally at the rate of convergence of the proximal gradient method (f convex in $F = f + g$)

$$\boxed{F(x_k) - F_* \;\leq\; \left(\frac{L_{\nabla f}}{2}\|x_0 - x_*\|^2\right)\frac{1}{k} \qquad \forall k \geq 0.}$$

This sublinear rate $(1/k)$ is better than the previous ones $(1/\sqrt{k})$ found before for the subgradient method (for minimizing f convex) and for the gradient method (for minimizing f nonconvex).

The WCC bound to reach $F(x_k) - F_* \leq \epsilon$ is then also better: $\mathcal{O}(\epsilon^{-1})$.

This applies also to $g = 0$, i.e., to the gradient method when f is convex.

In the strongly convex case $(\mu_f > 0)$, $(B3)$ gives also $(F(x_{k+1}) \geq F_*$ and $\alpha = \frac{1}{L_{\nabla f}})$

$$0 \;\leq\; \frac{L_{\nabla f}}{2}\left((1 - \mu_f/L_{\nabla f})\|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2\right)$$

$$\Updownarrow$$

$$\|x_{k+1} - x_*\|^2 \;\leq\; \left(1 - \frac{\mu_f}{L_{\nabla f}}\right)\|x_k - x_*\|^2$$

from which we conclude right away that $\|x_{k+1} - x_*\| \leq \|x_k - x_*\|$, i.e., the distance to the optimal set $X_* = \{x_*\}$ does not increase.

Moreover, we obtain the following rate of convergence for the proximal gradient method (f strongly convex in $F = f + g$)

$$\boxed{\|x_k - x_*\| \;\leq\; \left(\underbrace{1 - \frac{\mu_f}{L_{\nabla f}}}_{<1}\right)^k \|x_0 - x_*\|^2 \qquad \forall k.}$$