

Optimization for Data Sciences
Lecture 2
 Following the Steepest Direction

Samuel Vaïter

1. Following the steepest direction

ANALYSE MATHÉMATIQUE. — *Méthode générale pour la résolution des systèmes d'équations simultanées*; par M. AUGUSTIN CAUCHY.

« Étant donné un système d'équations simultanées qu'il s'agit de résoudre, on commence ordinairement par les réduire à une seule, à l'aide d'éliminations successives, sauf à résoudre définitivement, s'il se peut, l'équation résultante. Mais il importe d'observer, 1^o que, dans un grand nombre de cas, l'élimination ne peut s'effectuer en aucune manière; 2^o que l'équation résultante est généralement très-compiquée, lors même que les équations données sont assez simples. Pour ces deux motifs, on conçoit qu'il serait

Figure 1.1: First mention of the gradient descent algorithm by Augustin Cauchy (1847). “Méthode Générale Pour La Résolution Des Systèmes d'équations Simultanées”. In: *C.R. Acad. Sci. Paris* 25, pp. 536–538.

Suppose that you want to solve the problem

$$\min_{x \in [-1,1]^d} f(x),$$

and you want to achieve a ε -precision on the objective function f , *i.e.*, you want to obtain an estimate $\hat{x} \in [-1,1]^d$ such that $f(\hat{x}) - f(x^*) \leq \varepsilon$, where x^* is a minimizer of f over $[-1,1]^d$. Since f is continuous, you cannot have access to all values $f(x)$ when x describes the constraints $[-1,1]^d$, but you can allow yourself multiples calls to the “oracle” $f(x)$ in order to achieve this precision. A naive way to do so is to consider a discretization of $[-1,1]$ of precision ε , that is

$$G_\varepsilon = \{k\varepsilon : k \in \{-\lfloor \varepsilon^{-1} \rfloor, \dots, \lfloor \varepsilon^{-1} \rfloor\}\},$$

and $G_\varepsilon^d = G_\varepsilon \times \dots \times G_\varepsilon$ its counterpart in dimension d .

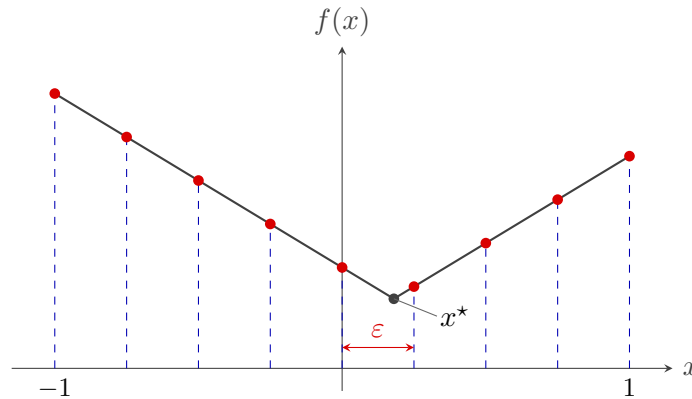


Figure 1.2: Discretization of a black-box function on $[-1,1]$. The red points are oracle evaluations, spaced by ε ; the minimizer x^* need not belong to the grid.

Proposition 1.1: Grid-search guarantee

Let $0 < \varepsilon \leq 1$ and let $f : [-1, 1]^d \rightarrow \mathbb{R}$ be L -Lipschitz, where $L > 0$. Let x^* be any minimizer of f on $[-1, 1]^d$, and let $\hat{x}$ be any minimizer of f over the finite grid G_ε^d . Then

$$0 \leq f(\hat{x}) - f(x^*) \leq L\sqrt{d}\varepsilon.$$

Thus, for a prescribed objective-value tolerance $\delta > 0$, choosing $\varepsilon \leq \min\{1, \delta/(L\sqrt{d})\}$ guarantees $f(\hat{x}) - f(x^*) \leq \delta$.

Proof. Lipschitz continuity implies continuity, so a minimizer x^* exists on the compact cube; a grid minimizer exists because the grid is finite and nonempty. For each coordinate, define

$$z_i = \varepsilon \operatorname{sign}(x_i^*) \left\lfloor \frac{|x_i^*|}{\varepsilon} \right\rfloor, \quad i = 1, \dots, d,$$

with $\operatorname{sign}(0) = 0$. Since $|x_i^*| \leq 1$, we have $z \in G_\varepsilon^d$ and $|z_i - x_i^*| < \varepsilon$ for every i . Consequently, $\|z - x^*\| \leq \sqrt{d}\varepsilon$. By global optimality of x^* , grid optimality of $\hat{x}$, and Lipschitz continuity,

$$0 \leq f(\hat{x}) - f(x^*) \leq f(z) - f(x^*) \leq L\|z - x^*\| \leq L\sqrt{d}\varepsilon.$$

□

This is a bound on objective values, rather than on $\|\hat{x} - x^*\|$. Lipschitz continuity alone cannot guarantee proximity to a chosen minimizer: for $f \equiv 0$, every grid point is a minimizer, including points far from $x^* = 0$.

The issue is that the size of G_ε^d grows exponentially fast with the dimension d . In dimension 1, one needs to perform $2\lfloor \varepsilon^{-1} \rfloor + 1$ evaluations of f , but in dimension d , one needs $(2\lfloor \varepsilon^{-1} \rfloor + 1)^d$! This quantity gets prohibitively large in a too fast manner: say you look at a very modest precision of $\varepsilon = 10^{-2}$, then in dimension $d = 1$, 201 computations of f are needed, whereas in dimension $d = 10$, you already need $201^{10} = 107636749520976961802001$ evaluations (about 1.08×10^{23}).

This is known as the curse of dimensionality, a term introduced by Bellman (1961, Chapter V) for optimal control and (over)used since then in the optimization, statistics and machine learning communities (Donoho, 2000). Summary: we need to be (slightly) more clever.

1.1. Descent direction

Consider the class of algorithms of the form

$$x^{(t+1)} = x^{(t)} + \eta^{(t)} u^{(t+1)},$$

how to “optimally choose” the direction $u^{(t+1)}$ such that $x^{(t+1)}$ is closer to a minimum of f ?

Definition 1.1

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$. A *descent direction* $u \in \mathbb{R}^d$ of f at $x \in \mathbb{R}^d$ is vector such that

$$\exists \varepsilon > 0, \forall \eta \in (0, \varepsilon), \quad f(x + \eta u) < f(x).$$

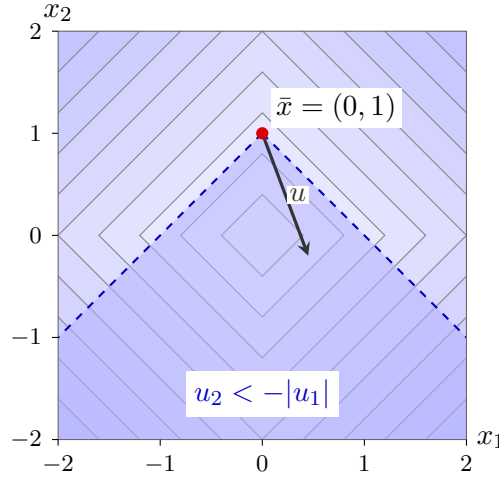


Figure 1.3: Descent cone of the nonsmooth function $f(x_1, x_2) = |x_1| + |x_2|$ at $\bar{x} = (0, 1)$. The level sets are diamonds. The shaded cone contains the strict descent directions $u_2 < -|u_1|$; its dashed boundary rays are excluded.

Observe that if f is differentiable, $\langle \nabla f(x), u \rangle < 0$ implies this property, but the converse need not hold.

For example, $f(z) = -z^2$ at $x = 0$ has the descent direction $u = 1$, but $f'(0)u = 0$.

The negative gradient $-\nabla f(\bar{x})$ is the *direction of steepest descent* at the point $\bar{x}$ when $\nabla f(\bar{x}) \neq 0$.

Proposition 1.2: The antigradient is the steepest direction

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a differentiable function at $\bar{x} \in \mathbb{R}^d$. Assume that $\nabla f(\bar{x}) \neq 0$. Then, the problem

$$\min_{\|u\|=1} \langle \nabla f(\bar{x}), u \rangle$$

has a unique solution $u^* = -\frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|}$.

Proof. Let $u \in \mathbb{R}^d$ such that $\|u\| = 1$. Consider the function $\phi_u : \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$\phi_u(t) = f(\bar{x} + tu).$$

The function ϕ_u is differentiable at 0 and its derivative reads $\phi'_u(0) = \langle \nabla f(\bar{x}), u \rangle$. Denoting θ the angle between $\nabla f(\bar{x})$ and u , we have that $\phi'_u(0) = \|\nabla f(\bar{x})\| \cos \theta$. Hence, minimizing $\langle \nabla f(\bar{x}), u \rangle$ is equivalent to finding the minimum of $\cos \theta$, that is $\theta = (2k+1)\pi$ for some $k \in \mathbb{Z}$. Thus, we have

$$u^* = -\frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|} \quad \text{and} \quad \langle \nabla f(\bar{x}), u^* \rangle = -\|\nabla f(\bar{x})\|.$$

□

1.2. The gradient descent (GD) algorithm

Proposition 1.2 gives birth to what is known as the *gradient descent algorithm* or *gradient descent method*.

Gradient Descent algorithm**Require:** Initialization $x^{(0)} \in \mathbb{R}^d$, step-size policy $\eta^{(t)} > 0$.1: **for** $t = 0, \dots$ **do**

$$x^{(t+1)} = x^{(t)} - \eta^{(t)} \nabla f(x^{(t)}). \quad (\text{GD})$$

The choice of the step-sizes $\eta^{(k)}$ is crucial, and there is several way to do it.

1. *Predetermined.* In this case, the sequence $(\eta^{(t)})$ is chosen beforehand, either with a constant step size $\eta^{(t)} = \eta \in \mathbb{R}_{>0}$ or with a given function of t , $\eta^{(t)} = g(t)$, e.g., $g(t) = \eta(t+1)^{-1/2}$ for some $\eta > 0$. For some classes of optimization problem, it is possible to provide guarantees depending on the choice of g .

2. *Oracle.* This is the “optimal” descent that a gradient descent method can produce, defined by

$$\eta_{\text{oracle}}^{(t)} \in \operatorname{argmin}_{\eta \geq 0} f(x^{(t)} - \eta \nabla f(x^{(t)})).$$

Remark that this choice is only theoretical since it involves a new optimization problem that may be nonsolvable in closed form.

This choice assumes that the line-search minimum is attained; this need not hold for a general f .

3. *Backtracking rule.* Start with a trial step-size and check whether the proposed gradient step decreases the objective sufficiently. If it does not, shrink the step-size by a fixed factor, for example divide it by two, and try again. Repeat until the test passes, then take the step. A common test is the Armijo condition

$$f(x^{(t)} - \eta \nabla f(x^{(t)})) \leq f(x^{(t)}) - c\eta \|\nabla f(x^{(t)})\|^2, \quad 0 < c < 1.$$

This adapts the step-size to the local behavior of the function without requiring its smoothness constant in advance.

1.3. Explicit analysis of the gradient descent: ordinary least-squares

All norms in the following gradient descent analysis are Euclidean.

Let $n, p \geq 1$, $A \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^n$. The least-square estimator is defined by the problem

$$\min_{x \in \mathbb{R}^p} f(x), \quad f(x) = \frac{1}{2n} \|Ax - y\|_2^2. \quad (1.1)$$

The objective f is twice differentiable, its gradient reads

$$\nabla f(x) = \frac{1}{n} A^\top (Ax - y) \in \mathbb{R}^p,$$

and its Hessian is constant on the whole space:

$$\nabla^2 f(x) = H = \frac{1}{n} A^\top A \in \mathbb{R}^{p \times p}.$$

Note that there is *at least* one minimizer to the problem (1.1). The first-order condition for a potential minimizer x^* is given by

$$Hx^* = \frac{1}{n} A^\top y.$$

A necessary, and sufficient condition, to have *exactly one* minimizer is that H is invertible.

The following lemma shows that, providing the knowledge of H , the t -iterates of the gradient descent with fixed step-size is computable in closed-form.

Lemma 1.1: Closed-form expression of the gradient descent step

For all $t \in \mathbb{N}$, we have

$$x^{(t)} - x^* = (I - \eta H)^t (x^{(0)} - x^*). \quad (1.2)$$

Proof. Using the expression of the gradient and (GD), we have for any minimizer x^*

$$\begin{aligned} x^{(t+1)} &= x^{(t)} - \frac{\eta}{n} A^\top (Ax^{(t)} - y) \\ &= x^{(t)} - \eta H(x^{(t)} - x^*). \end{aligned}$$

Subtracting x^* from both side gives

$$x^{(t+1)} - x^* = (I - \eta H)(x^{(t)} - x^*).$$

This identity is linear recursion, and unrolling it leads to the result. $\square$

1.3.1. Convergence in norm

We denote by μ the smallest eigenvalue of H , L its largest. If $L = 0$, then $A = 0$, f is constant and the GD iterates are stationary. In what follows, assume $L > 0$, and let $\kappa = L/\mu$ when $\mu > 0$, and $\kappa = +\infty$ when $\mu = 0$. We have $0 \leq \mu \leq L$ and $\kappa \geq 1$.

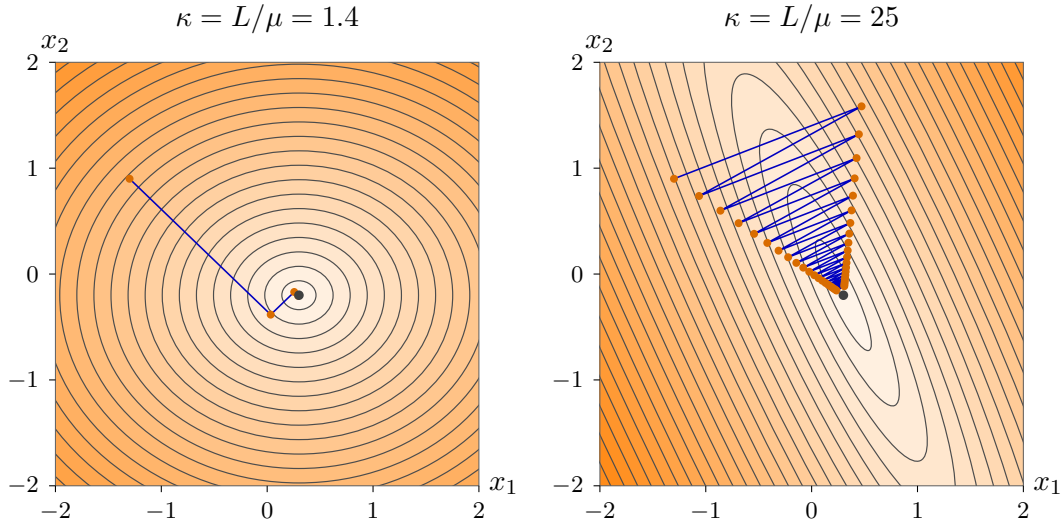


Figure 1.4: Gradient descent on two least-squares quadratics with the same initialization and minimizer, using $\eta = 2/(\mu + L)$. A small condition number gives fast convergence; a large condition number produces a slow zigzag along an elongated valley.

Using (1.2), we have

$$\|x^{(t)} - x^*\|^2 = \langle x^{(0)} - x^*, (I - \eta H)^{2t} (x^{(0)} - x^*) \rangle.$$

Remark that the eigenvalues of $(I - \eta H)^{2t}$ are exactly given by $(1 - \eta\lambda)^{2t}$ where λ is an eigenvalue of H . Since all eigenvalues of H are contained in $[\mu, L]$, one also can bound any $\rho \in \text{eigen}((I - \eta H)^{2t})$ by

$$|\rho| \leq \left(\max_{\lambda \in [\mu, L]} |1 - \eta\lambda| \right)^{2t}.$$

Suppose that there is multiple minimizers. In this case, H is not invertible and the smallest eigenvalue μ is equal to 0. Hence, there exists $v \neq 0$ such that $(I - \eta H)v = v$. Choose x^* a minimizer. Using $x^{(0)} = v + x^*$, we have

$$\|x^{(t)} - x^*\|^2 = \langle v + x^* - x^*, (I - \eta H)^{2t} (v + x^* - x^*) \rangle = \langle v, (I - \eta H)^{2t} v \rangle = \|v\|_2^2.$$

Hence, the method does not necessarily converge to this chosen x^* : here $x^{(t)} = x^{(0)}$ is already a minimizer. For $0 < \eta < 2/L$, the iterates do converge to a minimizer, depending on the initialization.

The situation is different when we assume that there is a unique minimizer.

Proposition 1.3: Convergence in norm

Assume that (1.1) has a unique minimizer x^* . Then, the gradient descent iterates $x^{(t)}$ with constant step-size $\eta^{(t)} = \eta = \frac{2}{\mu+L}$ converges linearly in norm

$$\|x^{(t)} - x^*\|^2 \leq \left(\frac{\kappa - 1}{\kappa + 1} \right)^{2t} \|x^{(0)} - x^*\|^2.$$

Proof. Since there is a unique minimizer x^* , we have $\mu > 0$. Observe that (using a spectral norm bound)

$$\langle x^{(0)} - x^*, (I - \eta H)^{2t} (x^{(0)} - x^*) \rangle \leq \left(\max_{\lambda \in [\mu, L]} |1 - \eta \lambda| \right)^{2t} \|x^{(0)} - x^*\|_2^2.$$

Observe that $\max_{\lambda \in [\mu, L]} |1 - \eta \lambda|$ is minimized for $\eta = \frac{2}{\mu+L}$, and value is $\frac{\kappa-1}{\kappa+1} \in [0, 1)$. Indeed,

$$\max_{\lambda \in [\mu, L]} |1 - \eta \lambda| = \max(|1 - \eta \mu|, |1 - \eta L|).$$

If $\mu = L$, this maximum is $|1 - \eta L|$, minimized at $\eta = 1/L$. For the following equivalences, assume $\mu < L$.

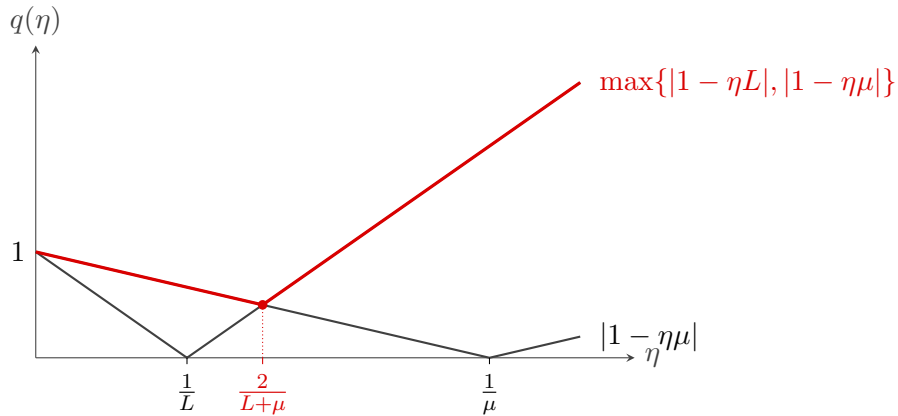


Figure 1.5: The bound $q(\eta) = \max\{|1 - \eta L|, |1 - \eta \mu|\}$ is minimized at the intersection, $\eta = 2/(L + \mu)$. The curves are drawn with $\mu = 1$ and $L = 3$.

Hence,

$$\min_{\eta > 0} \max_{\lambda \in [\mu, L]} |1 - \eta \lambda| = \min_{\eta > 0} \max(|1 - \eta \mu|, |1 - \eta L|).$$

Moreover, we have the following set of equivalence

$$\begin{aligned} |1 - \eta L| &\leq |1 - \eta \mu| \\ (1 - \eta L)^2 &\leq (1 - \eta \mu)^2 \\ \eta L^2 - 2L &\leq \eta \mu^2 - 2\mu \\ \eta &\leq \frac{2}{L + \mu}. \end{aligned}$$

This minima is achieved when the two curve $|1 - \eta L|$ and $|1 - \eta \mu|$ intersect. $\square$

Choice independant from the strong-convexity constant μ : It is possible to chose a smaller or equal stepsize (hence no faster) by taking the value $\eta = 1/L$. In this case, we get

$$\max_{\lambda \in [\mu, L]} |1 - \eta\lambda| = 1 - \frac{\mu}{L} = 1 - \frac{1}{\kappa}.$$

In both cases, we obtained a linear (or geometric, or exponential depending on the community) rate

$$\|x^{(t)} - x^*\|^2 \leq c^{2t} \|x^{(0)} - x^*\|^2,$$

where $c = \frac{\kappa-1}{\kappa+1}$ or $c = 1 - \frac{1}{\kappa}$ depending on the choice of the step-size.

It is possible to also speaks in term of iteration complexity. Assuming the choice of $\eta = 1/L$, observe that

$$\left(1 - \frac{1}{\kappa}\right)^{2t} \leq \exp\left(-\frac{1}{\kappa}\right)^{2t} = \exp\left(-\frac{2t}{\kappa}\right).$$

Thus, to obtain a fraction $\varepsilon \|x^{(0)} - x^*\|^2$ for $0 < \varepsilon < 1$, it is sufficient to perform t iterations with

$$t \geq \left\lceil \frac{\kappa}{2} \log \frac{1}{\varepsilon} \right\rceil.$$

1.3.2. Convergence in value

As we say, if $\mu = 0$, the iterates need not converge to a chosen minimizer, although they converge to a minimizer for $0 < \eta < 2/L$. But the story is different for the convergence in value.

Proposition 1.4: Convergence in value

Let x^* any solution of (1.1), assume $L > 0$, and let $t \geq 1$. The gradient descent iterates $x^{(t)}$ with constant step-size $\eta^{(t)} = \eta = \frac{1}{L}$ converges with a $O(1/t)$ rate

$$f(x^{(t)}) - f(x^*) \leq \frac{1}{4t\eta} \|x^{(0)} - x^*\|_2^2.$$

If moreover, (1.1) has a unique solution, the convergence is linear:

$$f(x^{(t)}) - f(x^*) \leq \left(1 - \frac{1}{\kappa}\right)^{2t} (f(x^{(0)}) - f(x^*)).$$

Proof. Observe that f has an *exact* Taylor expansion of order 2. Indeed, for any $x \in \mathbb{R}^p$ and any minimizer x^* :

$$\begin{aligned} f(x) - f(x^*) &= \langle \nabla f(x^*), x - x^* \rangle + \frac{1}{2} \langle x - x^*, H(x - x^*) \rangle \\ &= \frac{1}{2} \langle x - x^*, H(x - x^*) \rangle. \end{aligned} \tag{1.3}$$

Using the (exact) Taylor expansion (1.3) of f , we have

$$f(x^{(t)}) - f(x^*) = \frac{1}{2} \langle x^{(t)} - x^*, H(x^{(t)} - x^*) \rangle.$$

Using (1.2), we have

$$f(x^{(t)}) - f(x^*) = \frac{1}{2} \langle (I - \eta H)^t (x^{(0)} - x^*), H(I - \eta H)^t (x^{(0)} - x^*) \rangle.$$

Since $(I - \eta H)$ is symmetric, we have

$$f(x^{(t)}) - f(x^*) = \frac{1}{2} \langle x^{(0)} - x^*, (I - \eta H)^{2t} H(x^{(0)} - x^*) \rangle.$$

Using the fact that for a symmetric positive semidefinite matrix B , we have $\langle Bw, w \rangle \leq \|B\|_{sp} \|w\|^2$, with $B = (I - \eta H)^{2t}$ and $w = H^{1/2}(x^{(0)} - x^*)$, and that B commutes with H , we have

$$f(x^{(t)}) - f(x^*) \leq \|(I - \eta H)^{2t}\| \frac{1}{2} \langle x^{(0)} - x^*, H(x^{(0)} - x^*) \rangle.$$

Using again (1.3), we have

$$f(x^{(t)}) - f(x^*) \leq \|(I - \eta H)^{2t}\| (f(x^{(0)}) - f(x^*)).$$

Case $\mu > 0$. Now, we have according to the previous section, the following linear rate.

$$f(x^{(t)}) - f(x^*) \leq \left(1 - \frac{1}{\kappa}\right)^{2t} (f(x^{(0)}) - f(x^*)).$$

Case $\mu = 0$. Let's study the eigenvalues of¹ $(I - \eta H)^{2t} H$. We keep our analysis restricted to $0 < \eta \leq \frac{1}{L}$.

$$\begin{aligned} |\lambda(1 - \eta\lambda)^{2t}| &\leq \lambda \exp(-\eta\lambda)^{2t} \\ &= \lambda \exp(-2t\eta\lambda) \\ &= \frac{1}{2t\eta} 2t\eta\lambda \exp(-2t\eta\lambda) && \text{insert the term } 2t\eta \\ &\leq \frac{1}{2t\eta} \sup_{\rho \geq 0} \rho \exp(-\rho) && \text{crude bound} \\ &= \frac{1}{2et\eta} && \text{maximum at } \rho = 1 \\ &\leq \frac{1}{4t\eta} && e > 2. \end{aligned}$$

Hence, we obtain the claimed $O(1/t)$ rate. $\square$

1.4. Convergence of gradient descent for nonconvex functions

Unfortunately, not every functions are quadratic, and we need to deep diver in the analysis of the gradient descent algorithm to understand its behavior on generic function. We are going to explore several class of functions that leads to different rates, either in norm or in objective values.

1.4.1. Coercive continuously differentiable function

The following proposition show a very weak result without any quantification of the *rate of convergence* of gradient descent.

Proposition 1.5: Convergence of (GD) for coercive C^1 functions

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a coercive function and assume moreover that it is continuously differentiable. For any initial guess $x^{(0)}$, the *oracle* GD iterates $x^{(t)}$ converge up to a subsequence towards a critical point of f .

Proof. (i) The value sequence is (strictly) decreasing (unless it reaches a critical point). We recall that the oracle GD chose the step-size $\eta^{(t)}$ as

$$\phi^{(t)}(\eta) \stackrel{\text{def.}}{=} f(x^{(t)} - \eta \nabla f(x^{(t)})), \quad \eta^{(t)} \in \underset{\eta \geq 0}{\operatorname{argmin}} \phi^{(t)}(\eta).$$

¹Note the presence of H after $(I - \eta H)^{2t}$.

If $\nabla f(x^{(t)}) \neq 0$, coercivity implies $\phi^{(t)}(\eta) \rightarrow +\infty$ as $\eta \rightarrow +\infty$, so the oracle step exists. If the gradient is zero, the line-search objective is constant and the iterate stays unchanged.

Hence, evaluating $\phi^{(t)}$ at $\eta^{(t)}$ and 0 leads to

$$f(x^{(t+1)}) = \phi^{(t)}(\eta^{(t)}) \leq \phi^{(t)}(0) = f(x^{(t)}).$$

Suppose that $f(x^{(t+1)}) = f(x^{(t)})$. Thus, 0 is a minimizer of $\phi^{(t)}$ and therefore its one-sided first-order condition reads $\frac{d}{d\eta}\phi^{(t)}(0) \geq 0$. Given $\eta \geq 0$, we have

$$\frac{d}{d\eta}\phi^{(t)}(\eta) = \langle -\nabla f(x^{(t)}), \nabla f(x^{(t)} - \eta \nabla f(x^{(t)})) \rangle.$$

Evaluating this expression at $\eta = 0$ gives us $-\|\nabla f(x^{(t)})\|^2 \geq 0$, hence $\nabla f(x^{(t)}) = 0$ that means $x^{(t)}$ is a critical point.

(ii) *The iterates sequence is converging (up to a subsequence).* Consider the set $S = \{x \in \mathbb{R}^d : f(x) \leq f(x^{(0)})\}$. Remark that:

- f is continuous, thus S is closed;
- f is coercive, thus S is bounded.

Hence, S is compact. Observe that for all $t \geq 0$, $x^{(t)} \in S$. Using BOLZANO–WEIERSTRASS theorem, we get that the sequence $(x^{(t)})$ has a convergent subsequence towards some x^* . Let us denote it by $(x^{(\psi(t))})_{t \geq 0}$ where $\psi : \mathbb{N} \rightarrow \mathbb{N}$ is increasing.

The value sequence is decreasing and bounded below, so it converges to ℓ . By continuity, $\ell = f(x^*)$.

(iii) *The limit of the iterates sequence is a critical point.* Suppose that x^* is not a critical point. Using (i), we now that x^{**} defined as $x^{**} = x^* - \eta^* \nabla f(x^*) \neq x^*$, where $\eta^* > 0$ is the oracle step, is such that $f(x^{**}) < f(x^*)$. Consider the sequence $y^{(t)} = x^{(\psi(t))}$ converging to x^* . Remark that, using the continuity of the map $z \mapsto z - \eta^* \nabla f(z)$ for a continuously differentiable function f , we have

$$\lim_{t \rightarrow +\infty} (y^{(t)} - \eta^* \nabla f(y^{(t)})) = x^* - \eta^* \nabla f(x^*) = x^{**}. \quad (1.4)$$

By definition of $\phi^{(t)}$, we also have for all $t \geq 0$ that

$$\phi^{(\psi(t))}(\eta^*) \geq \phi^{(\psi(t))}(\eta^{(\psi(t))}) \geq \ell = f(x^*).$$

Hence,

$$f(y^{(t)} - \eta^* \nabla f(y^{(t)})) \geq f(x^*)$$

Using (1.4), we get that $f(x^{**}) \geq f(x^*)$, that is a contradiction. In conclusion, x^* is a critical point.

The same argument applies to any convergent subsequence. □

Remark that if in addition f has a unique critical point, which is then its unique global minimizer, then $x^{(t)}$ converges to this global minimizer.

Corollary 1.1

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a coercive, continuously differentiable function with a unique critical point x^* . For any initial guess $x^{(0)}$, the *oracle* GD iterates $x^{(t)}$ converge towards x^*

Proof. See Exercise 2. □

1.4.2. Smooth function bounded from below

Consider a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ that is L -smooth, but not necessarily convex. In addition, we suppose that f has full domain and is bounded below on $\mathbb{R}^d$. The following proposition gives a $O(1/\sqrt{T})$ rate for the 1st-order optimality using a constant step-size. A similar proof is doable for the oracle or the backtracking gradient descent.

Proposition 1.6: Convergence of (GD) for L -smooth functions

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a L -smooth function with $L > 0$, bounded from below by $\bar{f} \in \mathbb{R}$. For any initial guess $x^{(0)}$, the GD iterates $x^{(t)}$ with step-size $\eta^{(t)} = \eta \in (0, 2/L)$ satisfy $\nabla f(x^{(t)}) \rightarrow 0$, and moreover

$$\min_{0 \leq t \leq T} \|\nabla f(x^{(t)})\| \leq \frac{1}{\sqrt{T+1}} \left(\omega^{-1} L (f(x^{(0)}) - \bar{f}) \right)^{1/2},$$

where $\omega = 2\alpha(1 - \alpha)$, $0 < \alpha < 1$, and $\eta = \frac{2\alpha}{L}$.

Proof. (i) Bound on one step of gradient descent. For $x \in \mathbb{R}^d$, let $y = x - \eta \nabla f(x)$. Using proposition A.2, we have that

$$\begin{aligned} f(y) &\leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 \\ &= f(x) + \langle \nabla f(x), -\eta \nabla f(x) \rangle + \frac{L}{2} \|\eta \nabla f(x)\|^2 \\ &= f(x) - \eta \|\nabla f(x)\|^2 + \frac{L\eta^2}{2} \|\nabla f(x)\|^2 \\ &= f(x) - \eta \left(1 - \frac{L\eta}{2}\right) \|\nabla f(x)\|^2. \end{aligned}$$

To ensure that $f(y) \leq f(x)$, we need to require that $\eta(1 - \frac{L\eta}{2}) \geq 0$ that is $0 \leq \eta \leq \frac{2}{L}$. We can parameterize η as $\eta = \frac{2\alpha}{L}$ with $\alpha \in (0, 1)$. Hence, we get the descent

$$f(x) - f(y) \geq \frac{2\alpha(1 - \alpha)}{L} \|\nabla f(x)\|^2. \quad (1.5)$$

(ii) Convergence of 1st-order optimality. Applying (1.5) to $x = x^{(t)}$ and $y = x^{(t+1)}$, we get that

$$f(x^{(t)}) - f(x^{(t+1)}) \geq \frac{2\alpha(1 - \alpha)}{L} \|\nabla f(x^{(t)})\|^2. \quad (1.6)$$

Summing inequality (1.6) for $t = 0 \dots T$, we obtain

$$\sum_{t=0}^T f(x^{(t)}) - f(x^{(t+1)}) \geq \frac{2\alpha(1 - \alpha)}{L} \sum_{t=0}^T \|\nabla f(x^{(t)})\|^2.$$

Observing that the l.h.s. telescopes, we have

$$\frac{2\alpha(1 - \alpha)}{L} \sum_{t=0}^T \|\nabla f(x^{(t)})\|^2 \leq f(x^{(0)}) - f(x^{(T+1)}) \leq f(x^{(0)}) - \bar{f}, \quad (1.7)$$

by definition of $\bar{f}$. Since (1.7) is true for all $T \in \mathbb{N}$, we get $\nabla f(x^{(t)})$ converges towards 0.

(iii) Rate of convergence. Using the fact that for all $t = 0 \dots T$, we have

$$\min_{0 \leq s \leq T} \|\nabla f(x^{(s)})\| \leq \|\nabla f(x^{(t)})\|$$

for each $0 \leq t \leq T$, we have from (1.7) that

$$\frac{2\alpha(1-\alpha)}{L}(T+1) \min_{0 \leq t \leq T} \|\nabla f(x^{(t)})\|^2 \leq f(x^{(0)}) - \bar{f},$$

Hence, the claimed result. $\square$

Note that we cannot say *anything* in this context about the rate of convergence of $f(x^{(t)})$ or $(x^{(t)})$! One can show that $\eta = \frac{1}{L}$ is the optimal step-size with respect to the bound in (1.5).

1.5. Convergence of gradient descent for convex functions

1.5.1. Convex L -smooth function

When f is supposed to be convex, we can have a rate of convergence in objective values.

Proposition 1.7: Rate of (GD) for convex L -smooth functions

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex L -smooth function with $L > 0$ and at least one minimizer. For any initial guess $x^{(0)}$, the GD iterates $x^{(t)}$ with step-size $\eta^{(t)} = \eta \in (0, \frac{2}{L})$ converge towards an optimal point x^* of f , and moreover

$$f(x^{(t)}) - f^* \leq \frac{2(f(x^{(0)}) - f^*)\|x^{(0)} - x^*\|^2}{t\eta(2 - L\eta)(f(x^{(0)}) - f^*) + 2\|x^{(0)} - x^*\|^2}.$$

If $x^{(0)}$ is a minimizer, the iterates are stationary and the first bound is understood as zero. Moreover, if $\eta = \frac{1}{L}$, then

$$f(x^{(t)}) - f^* \leq \frac{2L\|x^{(0)} - x^*\|^2}{t + 4}.$$

Proof. Let x^* a minimizer of f .

If $f(x^{(0)}) = f(x^*)$, the iterates are stationary and the bounds are understood as zero. Otherwise $r^{(0)} > 0$. In the following divisions, assume $\delta^{(t+1)} > 0$; if an iterate is a minimizer, all later iterates stay there and the claimed bounds are immediate.

We consider $r^{(t)} = \|x^{(t)} - x^*\|$ the lack of optimality². We have

$$\begin{aligned} (r^{(t+1)})^2 &= \|x^{(t)} - x^* - \eta \nabla f(x^{(t)})\|^2 \\ &= (r^{(t)})^2 - 2\eta \langle \nabla f(x^{(t)}), x^{(t)} - x^* \rangle + \eta^2 \|\nabla f(x^{(t)})\|^2 \end{aligned}$$

Using the co-coercivity of the gradient (proposition A.3), we have that

$$\langle \nabla f(x^{(t)}) - \underbrace{\nabla f(x^*)}_{=0}, x^{(t)} - x^* \rangle \geq \frac{1}{L} \|\nabla f(x^{(t)}) - \underbrace{\nabla f(x^*)}_{=0}\|^2.$$

Thus,

$$(r^{(t+1)})^2 \leq (r^{(t)})^2 - \eta \left(\frac{2}{L} - \eta \right) \|\nabla f(x^{(t)})\|^2.$$

In particular, we get that $r^{(t)} \leq r^{(0)}$ for all $t \geq 0$.

Using the smoothness of f , proposition A.2 gives us the bound

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

²Note that we do not assume that $x^{(t)}$ converges towards this specific x^* .

Using the gradient descent update, we have that

$$f(x^{(t+1)}) \leq f(x^{(t)}) - \eta(1 - \frac{L\eta}{2}) \|\nabla f(x^{(t)})\|^2.$$

Now, let $\delta^{(t)} = f(x^{(t)}) - f(x^*)$. Using the differential characterization of convexity (proposition A.1), we have

$$\delta^{(t)} \leq \langle \nabla f(x^{(t)}), x^{(t)} - x^* \rangle.$$

Using CAUCHY-SCHWARZ inequality, we have

$$\delta^{(t)} \leq \|\nabla f(x^{(t)})\| \|x^{(t)} - x^*\| = r^{(t)} \|\nabla f(x^{(t)})\| \leq r^{(0)} \|\nabla f(x^{(t)})\|.$$

Hence, we have

$$\delta^{(t+1)} \leq \delta^{(t)} - \frac{q}{(r^{(0)})^2} (\delta^{(t)})^2,$$

where $q = \eta(1 - \frac{L\eta}{2})$. Multiplying both side by $\frac{1}{\delta^{(t)}\delta^{(t+1)}} > 0$, we get that

$$\frac{1}{\delta^{(t)}} \leq \frac{1}{\delta^{(t+1)}} - \frac{q}{(r^{(0)})^2} \frac{\delta^{(t)}}{\delta^{(t+1)}}.$$

Since $(f(x^{(t)}))$ is nonincreasing, we have $\frac{\delta^{(t)}}{\delta^{(t+1)}} \geq 1$, hence

$$\frac{1}{\delta^{(t+1)}} \geq \frac{1}{\delta^{(t)}} + \frac{q}{(r^{(0)})^2}.$$

Summing T inequalities of this type, we obtain

$$\frac{1}{\delta^{(T)}} \geq \frac{1}{\delta^{(0)}} + T \frac{q}{(r^{(0)})^2}.$$

We conclude using a bit of computation:

$$\delta^{(T)} \leq \left(\frac{1}{\delta^{(0)}} + T \frac{q}{(r^{(0)})^2} \right)^{-1} = \left(\frac{(r^{(0)})^2 + Tq\delta^{(0)}}{\delta^{(0)}(r^{(0)})^2} \right)^{-1} = \frac{\delta^{(0)}(r^{(0)})^2}{(r^{(0)})^2 + Tq\delta^{(0)}}$$

Replacing q , $r^{(0)}$ and $\delta^{(0)}$ by their expression gives the result.

Maximizing the descent $\eta(2 - L\eta)$ gives the optimal step-size $\eta = \frac{1}{L}$. Injecting this value, we get that

$$\delta^{(T)} \leq \frac{2L\delta^{(0)}(r^{(0)})^2}{2L(r^{(0)})^2 + T\delta^{(0)}}.$$

Using the smoothness (proposition A.2), we have

$$f(x^{(0)}) \leq f^* + \langle \nabla f(x^*), x^{(0)} - x^* \rangle + \frac{L}{2}(r^{(0)})^2 = f^* + \frac{L}{2}(r^{(0)})^2$$

Thus,

$$2L(r^{(0)})^2 + T\delta^{(0)} \geq (4 + T)\delta^{(0)},$$

which is the claimed result.

Finally, the nonincreasing distances to any minimizer make the iterates bounded. A convergent subsequence exists, and the value bound and continuity show that its limit $\bar{x}$ is a minimizer. Since $\|x^{(t)} - \bar{x}\|$ is nonincreasing and tends to zero along that subsequence, the whole sequence converges to $\bar{x}$.

□

1.5.2. Strongly convex L -smooth function

We now prove the linear convergence of (GD) when dealing with strongly convex functions.

Proposition 1.8: Rate of (GD) for strongly convex smooth functions

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a μ -strongly convex L -smooth function with $0 < \mu \leq L$. For any initial guess $x^{(0)}$, the GD iterates $x^{(t)}$ with step-size $\eta^{(t)} = \eta \in (0, \frac{2}{\mu+L}]$ converge towards an optimal point of f , and moreover

$$\|x^{(t)} - x^*\|^2 \leq \left(1 - \frac{2\eta\mu L}{\mu + L}\right)^t \|x^{(0)} - x^*\|^2.$$

Moreover, if $\eta = \frac{2}{\mu+L}$, then

$$\begin{aligned} \|x^{(t)} - x^*\|^2 &\leq \left(\frac{K_f - 1}{K_f + 1}\right)^{2t} \|x^{(0)} - x^*\|^2 \\ f(x^{(t)}) - f^* &\leq \frac{L}{2} \left(\frac{K_f - 1}{K_f + 1}\right)^{2t} \|x^{(0)} - x^*\|^2, \end{aligned}$$

where $K_f = L/\mu$.

Proof. Let $r^{(t)} = \|x^{(t)} - x^*\|$. We start like the proof of proposition 1.7:

$$\begin{aligned} (r^{(t+1)})^2 &= \|x^{(t)} - x^* - \eta \nabla f(x^{(t)})\|^2 \\ &= (r^{(t)})^2 - 2\eta \langle \nabla f(x^{(t)}), x^{(t)} - x^* \rangle + \eta^2 \|\nabla f(x^{(t)})\|^2 \end{aligned}$$

Using proposition A.5, we have

$$\langle \nabla f(x^{(t)}) - \nabla f(x^*), x^{(t)} - x^* \rangle \geq \frac{\mu L}{\mu + L} \|x^{(t)} - x^*\|^2 + \frac{1}{\mu + L} \|\nabla f(x^{(t)}) - \nabla f(x^*)\|^2,$$

that is

$$\langle \nabla f(x^{(t)}) - \nabla f(x^*), x^{(t)} - x^* \rangle \geq \frac{\mu L}{\mu + L} (r^{(t)})^2 + \frac{1}{\mu + L} \|\nabla f(x^{(t)})\|^2.$$

Hence,

$$(r^{(t+1)})^2 \leq \left(1 - \frac{2\eta\mu L}{\mu + L}\right) (r^{(t)})^2 + \eta \left(\eta - \frac{2}{\mu + L}\right) \|\nabla f(x^{(t)})\|^2.$$

Since $\eta - \frac{2}{\mu+L} \leq 0$, we can drop it and we obtain the recursion $(r^{(t+1)})^2 \leq (1 - q)(r^{(t)})^2$ with $q = \frac{2\eta\mu L}{\mu+L}$, that proves the first claim (linear rate).

Let $\eta = \frac{2}{\mu+L}$. We have

$$1 - q = 1 - \frac{4\mu L}{(\mu + L)^2} = \frac{(L - \mu)^2}{(L + \mu)^2} = \left(\frac{K_f - 1}{K_f + 1}\right)^2.$$

The last inequality is yet another use of proposition A.2. □

TODO Polyak assumption

1.6. Introduction to lower bounds

All the bounds up to now that we described and proved were *upper* bounds of the form $\|x^{(t)} - x^*\| \leq \alpha(t)$ or $f(x^{(t)}) - f(x^*) \leq \alpha(t)$. But what about finding the reverse-side inequality $\|x^{(t)} - x^*\| \geq \beta(t)$? Said otherwise, what can we achieve with a “gradient-descent-like” algorithm? To formalize this notion,

Assumption 1.1: First-order method

We assume that a first-order method is given by a sequence $x^{(t)}$ such that

$$x^{(t)} \in x^{(0)} + \text{Span}\{\nabla f(x^{(0)}), \dots, \nabla f(x^{(t-1)})\}.$$

We shall note that one can think of a more general way to define first-order methods, but for the sake of the results we aim to prove, such level of generality is enough.

With this assumption in mind, how to design a function adversarial to these type of schemes? The idea is to find a function such that the gradient at step $t - 1$ gives minimal information, *i.e.*, it has a minimal nonzero partial derivatives. A way to define such function is to “stack” quadratic functions with increasing dependencies between variables:

$$f_k^{L,\mu}(x) = \frac{L-\mu}{8} \left((x_1 - 1)^2 + \sum_{i=1}^{k-1} (x_{i+1} - x_i)^2 + x_k^2 \right) + \frac{\mu}{2} \|x\|^2, \quad (1.8)$$

where $0 \leq \mu < L$ and $1 \leq k \leq d$.

$$\frac{\partial f_k^{L,\mu}}{\partial x_i}(x) = \mu x_i + \frac{L-\mu}{4} \begin{cases} -x_2 + 2x_1 - 1 & \text{if } i = 1 < k, \\ -x_{i+1} + 2x_i - x_{i-1} & \text{if } 2 \leq i < k, \\ 2x_k - x_{k-1} & \text{if } i = k > 1, \\ 2x_1 - 1 & \text{if } i = k = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Set $f = f_k^{L,\mu}$. Observe that if we start from $x^{(0)} = 0$, then

$$x^{(1)} = x^{(0)} - \eta \nabla f(0) = +\eta \frac{L-\mu}{4} e_1 \in \mathbb{R}e_1,$$

that is only the first coordinate is updated after one iteration. What happens now that we have access to $x^{(0)}$, $\nabla f_k(x^{(0)})$ and $\nabla f_k(x^{(1)})$? An algorithm satisfying Assumption 1.1, we look at

$$x^{(2)} = x^{(0)} + \alpha \nabla f(x^{(0)}) + \beta \nabla f(x^{(1)}).$$

One can check that for any (α, β) , $x^{(2)} \in \mathbb{R}e_1 + \mathbb{R}e_2$, and by an easy induction, we have $x^{(t)} \in \sum_{k=1}^t \mathbb{R}e_k$: any first-order methods will only be able to update at most one new coordinate at each iteration. We are going to prove the following result.

Theorem 1.1: Lower-bound for smooth convex optimization

For any $d \geq 2$, $x^{(0)} \in \mathbb{R}^d$, $L > 0$, $t \in \mathbb{N}$, $0 \leq t \leq (d-1)/2$, there exists a convex function f that is C^∞ and L -smooth such that any sequences satisfying Assumption 1.1 is such that

$$f(x^{(t)}) - f(x^*) \geq \frac{3L\|x^{(0)} - x^*\|^2}{32(t+1)^2} \quad (1.9)$$

$$(1.10)$$

where x^* is a minimizer of f .

Remark that the rate $1/t^2$ is not attained by the gradient descent! We will see a latter lecture strategies to achieve this rate. We also have a lower bound for the class of strongly convex functions.

Theorem 1.2: Lower-bound for smooth strongly convex optimization

For any $d \geq 2$, $x^{(0)} \in \mathbb{R}^d$, $0 < \mu < L$, there exists a μ -strongly convex function f that is C^∞ and L -smooth such that any sequences satisfying Assumption 1.1 is such that for all $t \in \mathbb{N}$ with $0 \leq t \leq (d-1)/2$, we have

$$\|x^{(t)} - x^*\|^2 \geq \frac{1}{8} \left(\frac{\sqrt{K_f} - 1}{\sqrt{K_f} + 1} \right)^{2t} \|x^{(0)} - x^*\|^2, \quad (1.11)$$

$$f(x^{(t)}) - f(x^*) \geq \frac{\mu}{16} \left(\frac{\sqrt{K_f} - 1}{\sqrt{K_f} + 1} \right)^{2t} \|x^{(0)} - x^*\|^2. \quad (1.12)$$

where x^* is the unique minimizer of f and $K_f = L/\mu$.

Note that it is common in the litterature to see Theorem 1.2 without the factor $\frac{1}{8}$. This is due to an artefact of proof since we prove this result in the finite dimensional case whereas Nesterov (2018) works in the infinite dimensional space $\ell^2(\mathbb{N})$. Before proving these important results due to (Nemirovski and Yudin, 1983), we are going to prove several lemmas.

Lemma 1.2: Minimizers of f_k

Let $d \geq 2$, $L > 0$, $0 \leq \mu < L$ and $1 \leq k \leq d$, then $f_k^{L,\mu}$ defined in (1.8) is a μ -strongly convex (eventually convex if $\mu = 0$) C^∞ -function such that its gradient is L -Lipschitz. If $\mu = 0$, it has a minimizer $x^{k,*}$ satisfying

$$x_i^{k,*} = \begin{cases} 1 - \frac{i}{k+1} & \text{if } 1 \leq i \leq k \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad f_k^{L,0}(x^{k,*}) = \frac{L}{8(k+1)}.$$

This minimizer is unique if $k = d$. If $k < d$, every minimizer has the displayed first k coordinates, while its remaining coordinates are arbitrary; the display chooses them to be zero.

If $\mu > 0$, it has a unique minimizer $x^{k,*}$ satisfying

$$x_i^{k,*} = \frac{s^{2(k+1)}}{s^{2(k+1)} - 1} s^{-i} + \frac{1}{1 - s^{2(k+1)}} s^i,$$

for $1 \leq i \leq k$, and $x_i^{k,*} = 0$ for $i > k$, where $s = \frac{\sqrt{K_f+1}}{\sqrt{K_f-1}}$.

Proof. We drop the exponents L, μ in the definition of $f_k = f_k^{L,\mu}$. The function f_k being a quadratic form, it is C^∞ and its partial derivatives read

$$\frac{\partial^2 f_k}{\partial x_i \partial x_j}(x) = \mu 1_{\{i=j\}} + \frac{L - \mu}{4} \begin{cases} 2 & \text{if } i = j \leq k \\ -1 & \text{if } j = i - 1 \text{ and } 1 < i \leq k \\ -1 & \text{if } j = i + 1 \text{ and } 1 \leq i < k \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the Hessian matrix is given (for any $x \in \mathbb{R}^d$) by

$$\nabla^2 f_k(x) = \mu \text{Id}_d + \frac{L - \mu}{4} L_k,$$

where L_k is a (thresholded) discrete Laplacian operator with Dirichlet boundary conditions

that is tridiagonal

$$L_k = \left(\begin{array}{ccc|ccc} 2 & -1 & 0 & & & \\ -1 & 2 & -1 & & & \\ & -1 & \ddots & \ddots & & \\ & & \ddots & \ddots & -1 & \\ & & & -1 & 2 & \\ \hline & & & 0_{d-k,k} & & 0_{d-k,d-k} \end{array} \right).$$

Observe that we have (since f_k is a quadratic form)

$$f_k(x) = \frac{1}{2} \langle \nabla^2 f_k(x) x, x \rangle - \frac{L - \mu}{4} x_1 + \frac{L - \mu}{8}.$$

Note that:

1. The Hessian is definite (resp. semi-definite) positive if $\mu > 0$ (resp. $\mu = 0$). Indeed,

$$\langle \nabla^2 f_k(x) h, h \rangle = \mu \|h\|^2 + \frac{L - \mu}{4} \langle L_k h, h \rangle.$$

Since $\langle L_k h, h \rangle = h_1^2 + \sum_{i=1}^{k-1} (h_{i+1} - h_i)^2 + h_k^2 \geq 0$ for any h , the result follows depending on the value of μ .

2. Since $(a - b)^2 \leq 2a^2 + 2b^2$, we have

$$\begin{aligned} h_1^2 + \sum_{i=1}^{k-1} (h_{i+1} - h_i)^2 + h_k^2 &\leq h_1^2 + \sum_{i=1}^{k-1} (2h_{i+1}^2 + 2h_i^2) + h_k^2 \\ &\leq 4 \sum_{i=1}^k h_i^2 \leq 4 \sum_{i=1}^d h_i^2 = 4\|h\|^2. \end{aligned}$$

Hence,

$$\langle \nabla^2 f_k(x) h, h \rangle \leq \mu \|h\|^2 + (L - \mu) \|h\|^2 = L \|h\|^2.$$

Thus, we have $\mu \text{Id} \preceq \nabla^2 f_k(x) \preceq L \text{Id}$.

Let us characterize a solution $x^{k,*}$, unique if $\mu > 0$ or $k = d$ of the minimization of f_k over $\mathbb{R}^d$. We aim to solve $\nabla f_k(x^{k,*}) = 0$ to find a critical point (which will be a minimum since we just proved that the Hessian is at least semidefinite positive), that is

$$\mu x^{k,*} + \frac{L - \mu}{4} L_k x^{k,*} - \frac{L - \mu}{4} e_1 = 0.$$

Projecting this relation on each coordinate $2 \leq i \leq k - 1$, we get that

$$-x_{i-1}^{k,*} + 2x_i^{k,*} - x_{i+1}^{k,*} = -\frac{4\mu}{L - \mu} x_i^{k,*},$$

which leads to

$$x_i^{k,*} = \frac{1}{2} \frac{L - \mu}{L + \mu} (x_{i+1}^{k,*} + x_{i-1}^{k,*}).$$

For the following two boundary equations, assume $k \geq 2$. If $k = 1$, the single equation gives $x_1^{1,*} = (L - \mu)/(2(L + \mu))$. The boundary notation below covers both cases.

Similarly, we have

$$x_1^{k,\star} = \frac{1}{2} \frac{L - \mu}{L + \mu} (x_2^{k,\star} + 1) \quad \text{and} \quad x_k^{k,\star} = \frac{1}{2} \frac{L - \mu}{L + \mu} x_{k-1}^{k,\star}.$$

Consider $y_0, \dots, y_{k+1}$ defined by $y_i = x_i^{k,\star}$ for $1 \leq i \leq k$ and $y_0 = 1$ and $y_{k+1} = 0$. We have the relation

$$y_i = \alpha(y_{i+1} + y_{i-1}) \quad \text{where} \quad \alpha = \frac{1}{2} \frac{L - \mu}{L + \mu} > 0.$$

We can rewrite it as the second-order linear recursion $y_{i+2} - \alpha^{-1}y_{i+1} + y_i = 0$. The associated trinomial is $P = X^2 - \alpha^{-1}X + 1 \in \mathbb{R}[X]$ whose discriminant is given by

$$\Delta = (-\alpha^{-1})^2 - 4 = 16 \frac{L\mu}{(L - \mu)^2}.$$

We distinguish two cases:

1. If $\mu = 0$, then the unique root is given by $r = 1$.
2. If $\mu > 0$, then the roots are given by

$$r = \frac{1}{2}(\alpha^{-1} - \sqrt{\Delta}) = \frac{\sqrt{\frac{L}{\mu}} - 1}{\sqrt{\frac{L}{\mu}} + 1} = \frac{\sqrt{K_f} - 1}{\sqrt{K_f} + 1}$$

$$s = \frac{1}{2}(\alpha^{-1} + \sqrt{\Delta}) = \frac{\sqrt{K_f} + 1}{\sqrt{K_f} - 1} = \frac{1}{r}.$$

Case $\mu = 0$. We have the affine relation $y_i = (a + bi)r$ with constraints $y_0 = a = 1$ and $y_{k+1} = a + b(k+1) = 0$. In turn, we have $y_i = 1 - \frac{i}{k+1}$ and thus

$$x_i^{k,\star} = \begin{cases} 1 - \frac{i}{k+1} & \text{if } 1 \leq i \leq k \\ 0 & \text{otherwise.} \end{cases}$$

The associated optimal value is given by

$$f_k(x^{k,\star}) = \frac{L}{8} \left(\left(-\frac{1}{k+1} \right)^2 + \sum_{i=1}^{k-1} \frac{1}{(k+1)^2} + \left(1 - \frac{k}{k+1} \right)^2 \right)$$

$$= \frac{L}{8} \frac{k+1}{(k+1)^2} = \frac{L}{8} \frac{1}{k+1}.$$

Case $\mu > 0$. The solution can be written as

$$y_i = ar^i + bs^i \quad \text{with} \quad \begin{cases} a + b & = 1 \\ ar^{k+1} + bs^{k+1} & = 0. \end{cases}$$

Thus, we have $b = 1 - a$, hence $\frac{a}{a-1} = s^{2(k+1)} > 0$, and in turn we have

$$a = \frac{s^{2(k+1)}}{s^{2(k+1)} - 1} \quad \text{and} \quad b = \frac{1}{1 - s^{2(k+1)}}.$$

Hence,

$$y_i = \frac{s^{2(k+1)}}{s^{2(k+1)} - 1} s^{-i} + \frac{1}{1 - s^{2(k+1)}} s^i.$$

□

Proof of Theorem 1.1. We restrict our attention to the case where $x_0 = 0$ w.l.o.g. Indeed, if $x_0 \neq 0$, we can set $x \mapsto \tilde{f}(x) = f(x + x_0)$ and the following proof carry on. Let $k = t$, so that $2k + 1 \leq d$, and set $f = f_{2k+1}^{L,0}$ on $\mathbb{R}^d$.

If $t = 0$, then $f(0) - f(x^*) = L/16$ and $\|x^*\|^2 = 1/4$, which gives the claimed bound. In what follows, assume $t = k \geq 1$.

Remark that

$$f(x^{(t)}) = f_{2k+1}^{L,0}(x^{(t)}) = f_t^{L,0}(x^{(t)}) \geq f_t^*.$$

Using Lemma 1.2, we have on one hand that $f_t^* = \frac{L}{8(t+1)}$, and then

$$f(x^{(k)}) - f(x^*) \geq \frac{L}{8(k+1)} - \frac{L}{16(k+1)} = \frac{L}{16(k+1)}.$$

On the other hand,

$$\begin{aligned} \|x^{2k+1,*} - x_0\|^2 &= \|x^{2k+1,*}\|^2 = \sum_{i=1}^{2k+1} (x^{2k+1,*})_i^2 \\ &= \sum_{i=1}^{2k+1} \left(1 - \frac{i}{2(k+1)}\right)^2 \\ &= \sum_{i=1}^{2k+1} 1 - \frac{2}{2(k+1)} \sum_{i=1}^{2k+1} i + \frac{1}{4(k+1)^2} \sum_{i=1}^{2k+1} i^2 \\ &= (2k+1) - \frac{1}{k+1} \frac{2(k+1)(2k+1)}{2} + \frac{1}{4(k+1)^2} \frac{2(k+1)(4k+3)(2k+1)}{6} \\ &= \frac{1}{3} \frac{(2k+1)(4k+3)}{4(k+1)} \\ &\leq \frac{2k+1}{3} \leq \frac{2}{3}(k+1). \end{aligned}$$

Thus,

$$\frac{f(x^{(t)}) - f(x^*)}{\|x^{2k+1,*} - x_0\|^2} \geq \frac{\frac{L}{16(k+1)}}{\frac{2}{3}(k+1)} = \frac{3L}{32(k+1)^2},$$

that proves (1.9). □

Proof of Theorem 1.2. The proof follows the same strategy as before, but we start with a bound on the iterates instead of the objective function. Assume that $x^{(0)} = 0$, otherwise let $\tilde{f} = f(\cdot + x_0)$. Let $k = \lfloor (d-1)/2 \rfloor$, $D = 2k+1 \leq d$, and $f = f_D^{L,\mu}$ on $\mathbb{R}^d$. We rewrite the coordinate of $x^{2k+1,*}$ as

$$x_i^{2k+1,*} = \frac{s^{4(k+1)}}{s^{4(k+1)} - 1} s^{-i} + \frac{1}{1 - s^{4(k+1)}} s^i = s^{-i} \left(1 - \frac{s^{2i} - 1}{s^{4(k+1)} - 1}\right).$$

On one hand, we have:

$$\|x^{(0)} - x^{2k+1,*}\|^2 = \sum_{i=1}^{2k+1} (x_i^{2k+1,*})^2 = \sum_{i=1}^{2k+1} s^{-2i} \left(1 - \frac{s^{2i} - 1}{s^{4(k+1)} - 1}\right)^2 \leq \sum_{i=1}^{2k+1} s^{-2i},$$

where we used that for all $1 \leq i \leq 2k+1$, we have

$$0 \leq 1 - \frac{s^{2i} - 1}{s^{4(k+1)} - 1} \leq 1.$$

Bounding the tail of the geometric sums, we obtain

$$\|x^{(0)} - x^{2k+1,*}\|^2 \leq 2 \sum_{i=1}^{k+1} s^{-2i}. \quad (1.13)$$

On the other hand, observe that for $0 \leq t \leq k$, the coordinates of $x^{(t)}$ after the first t are zero. Let $a = \log s > 0$. We rewrite the coordinate of $x^{D,*}$ as

$$x_i^{D,*} = \frac{\sinh((D+1-i)a)}{\sinh((D+1)a)}, \quad 1 \leq i \leq D.$$

Thus, setting $m = D - t$, we have

$$\|x^{(t)} - x^{D,*}\|^2 \geq \frac{\sum_{j=1}^m \sinh^2(ja)}{\sinh^2((D+1)a)}, \quad \|x^{(0)} - x^{D,*}\|^2 = \frac{\sum_{j=1}^D \sinh^2(ja)}{\sinh^2((D+1)a)}.$$

We now compare these two sums. For each $1 \leq j \leq D$, let $u = \lceil jm/D \rceil$. Since $D \geq 2t + 1$, we have $D/m < 2$, $1 \leq u \leq m$, $u \leq j$, and $j - u \leq t$; each value of u appears at most twice. Concavity of $v \mapsto 1 - e^{-2av}$, which is zero at $v = 0$, gives

$$\frac{\sinh(ja)}{\sinh(ua)} = e^{(j-u)a} \frac{1 - e^{-2ja}}{1 - e^{-2ua}} \leq e^{ta} \frac{j}{u} \leq 2s^t.$$

Hence,

$$\sum_{j=1}^D \sinh^2(ja) \leq 8s^{2t} \sum_{u=1}^m \sinh^2(ua).$$

Combining it with the preceding expressions, we have

$$\|x^{(t)} - x^{D,*}\|^2 \geq \frac{1}{8} s^{-2t} \|x^{(0)} - x^{D,*}\|^2 = \frac{1}{8} \left(\frac{\sqrt{K_f} - 1}{\sqrt{K_f} + 1} \right)^{2t} \|x^{(0)} - x^{D,*}\|^2,$$

proving (1.11). The value bound (1.12) is obtained by applying (A.2) to this bound. $\square$

1.7. An invitation to flow

The joint study of ordinary differential equations and iterative algorithms is very fruitful. Imagine that the iterate $x^{(t)}$ can be viewed as “snapshot” $x^{(t)} = x(t\eta)$ of a smooth curve $x : [0, +\infty) \rightarrow \mathbb{R}^d$, for some learning rate $\eta > 0$. Let $\tau = t\eta$, observe that

$$x(\tau + \eta) = x^{(t+1)} = x^{(t)} - \eta \nabla f(x^{(t)}) = x(\tau) - \eta \nabla f(x(\tau)).$$

Reorganizing the terms, we get that

$$\frac{1}{\eta} (x(\tau + \eta) - x(\tau)) = -\nabla f(x(\tau)).$$

Assuming enough regularity, we can let η goes to 0, and we obtain the ODE

$$\dot{x}(\tau) = -\nabla f(x(\tau)),$$

where $\dot{x}(\tau) = \frac{dx}{d\tau}(\tau)$ is the derivative of x along the time. The Cauchy problem

$$\begin{cases} \dot{x}(\tau) = -\nabla f(x(\tau)) & \text{if } \tau \geq 0 \\ x(0) = x^{(0)}, \end{cases} \quad (\text{GF})$$

is known as the *gradient flow* of f . One can thus interpret the step-size (or learning rate) η as the discretization precision of the gradient flow. It is possible to derive a similar convergence result for (GF) as for the gradient descent.

Theorem 1.3: Convergence of the gradient flow

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex, continuously differentiable, and have at least one minimizer. Then, for any initial point $x^{(0)}$, the dynamics $\tau \mapsto x(\tau)$ converges towards an optimal point x^* of f , and moreover, for all $\tau > 0$, one has

$$f(x(\tau)) - f(x^*) \leq \frac{\|x^{(0)} - x^*\|^2}{2\tau}.$$

Proof. Let z a minimizer of f . By convexity, we have

$$\frac{d}{d\tau} \frac{1}{2} \|x(\tau) - z\|^2 = -\langle \nabla f(x(\tau)), x(\tau) - z \rangle \leq -(f(x(\tau)) - f(z)) \leq 0.$$

The continuous vector field gives a local solution. Its distance to z is nonincreasing, so it stays bounded and extends to all $\tau \geq 0$. Monotonicity of ∇f makes the distance between two solutions with the same starting point nonincreasing, so the solution is unique. Moreover,

$$\frac{d}{d\tau} f(x(\tau)) = -\|\nabla f(x(\tau))\|^2 \leq 0.$$

Integrating the first inequality and using this monotonicity, we obtain

$$\tau(f(x(\tau)) - f(z)) \leq \int_0^\tau (f(x(s)) - f(z)) ds \leq \frac{1}{2} \|x^{(0)} - z\|^2.$$

Hence, the claimed value bound holds for any minimizer z . Since the trajectory is bounded, it has a limit point $\bar{x}$ along a sequence of times tending to infinity. The value bound and continuity show that $\bar{x}$ is a minimizer. Its distance to the trajectory is nonincreasing and tends to zero along that sequence, so $x(\tau)$ converges to $\bar{x}$. Taking $z = \bar{x} = x^*$ gives the displayed bound. □

1.8. Complements**1.8.1. Optimization for multinomial regression**

We study here a method for classification with K classes called the *multinomial logistic regression* or *softmax classification*. The objective function reads for $W \in \mathbb{R}^{p \times K}$,

$$f(W) = - \sum_{i=1}^n \sum_{k=1}^K \mathbf{1}_{y_i=k} \log(\hat{p}(y_i = k \mid x_i)) + \lambda \|W\|_{\text{Fro}}^2, \quad (1.14)$$

where

$$\hat{p}(y_i = k \mid x_i) = \frac{\exp(x_i^\top W_k)}{\sum_{l=1}^K \exp(x_i^\top W_l)}.$$

Here $x_i \in \mathbb{R}^p$, $y_i \in \{1, \dots, K\}$, W_k denotes the k -th column of W , and $\lambda \geq 0$. The *softmax* function $\sigma : \mathbb{R}^K \rightarrow \mathbb{R}^K$ is defined as

$$\sigma(x) = (\sigma_1(x), \dots, \sigma_K(x))^\top \quad \text{where} \quad \sigma_i(x) = \frac{e^{x_i}}{\sum_{j=1}^K e^{x_j}}.$$

We also introduce the *negative log-softmax* functions $\psi_i : \mathbb{R}^K \rightarrow \mathbb{R}$ defined as

$$\psi_k(x) = -\log \sigma_k(x).$$

Thus, our loss is expressed as

$$f(W) = \sum_{i=1}^n \sum_{k=1}^K \mathbf{1}_{y_i=k} \psi_k(W^\top x_i) + \lambda \|W\|_{\text{Fro}}^2,$$

Lemma 1.3: Jacobian of the softmax

Let $x \in \mathbb{R}^K$,

$$\text{Jac}_\sigma(x) = \text{diag}(\sigma(x)) - \sigma(x)\sigma(x)^\top.$$

Proof. Observe that for $j \in \{1, \dots, K\}$, we have

$$\frac{\partial \left(\sum_{k=1}^K e^{x_k} \right)^{-1}}{\partial x_j} = -\frac{1}{\left(\sum_{k=1}^K e^{x_k} \right)^2} \frac{\partial \left(\sum_{k=1}^K e^{x_k} \right)}{\partial x_j} = -\frac{e^{x_j}}{\left(\sum_{k=1}^K e^{x_k} \right)^2} = -\frac{1}{\sum_{k=1}^K e^{x_k}} \sigma_j(x).$$

Now, take $i, j \in \{1, \dots, K\}$. Then, two cases occurs:

If $i \neq j$. We have

$$\frac{\partial \sigma_i}{\partial x_j}(x) = e^{x_i} \frac{\partial \left(\sum_{k=1}^K e^{x_k} \right)^{-1}}{\partial x_j} = -\sigma_i(x) \sigma_j(x).$$

If $i = j$.

$$\frac{\partial \sigma_j}{\partial x_j}(x) = e^{x_j} \frac{\partial \left(\sum_{k=1}^K e^{x_k} \right)^{-1}}{\partial x_j} + e^{x_j} \left(\sum_{k=1}^K e^{x_k} \right)^{-1} = \sigma_j(x)(1 - \sigma_j(x)).$$

□

Lemma 1.4: Gradient of the negative logsoftmax

Let $i \in \{1, \dots, K\}$ and $x \in \mathbb{R}^K$. Then,

$$\nabla \psi_i(x) = \sigma(x) - e_i \quad \text{and} \quad \nabla^2 \psi_i(x) = \text{Jac}_\sigma(x).$$

Proof. Let $k \in \{1, \dots, K\}$. Then,

$$\frac{\partial \psi_i}{\partial x_k}(x) = -\frac{1}{\sigma_i(x)} \frac{\partial \sigma_i}{\partial x_k}(x) = \begin{cases} \sigma_k(x) & \text{if } i \neq k \\ \sigma_k(x) - 1 & \text{if } i = k. \end{cases}$$

Differentiating $\nabla \psi_i(x) = \sigma(x) - e_i$ gives $\nabla^2 \psi_i(x) = \text{Jac}_\sigma(x)$.

□

Lemma 1.5: Convexity of the negative logsoftmax

For all $k \in \{1, \dots, K\}$, ψ_k is convex.

Proof. Using the preceding lemma, we have for any $v \in \mathbb{R}^K$

$$v^\top \nabla^2 \psi_k(x) v = \sum_{j=1}^K \sigma_j(x) v_j^2 - \left(\sum_{j=1}^K \sigma_j(x) v_j \right)^2 = \sum_{j=1}^K \sigma_j(x) (v_j - \bar{v})^2 \geq 0,$$

where $\bar{v} = \sum_{j=1}^K \sigma_j(x) v_j$, since $\sum_j \sigma_j(x) = 1$. Hence, ψ_k is convex.

□

1.9. Exercises

Exercise 1 Consider the problem

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{2} \|Ax - b\|^2,$$

where $A \in \mathbb{R}^{n \times d}$, $n, d \geq 1$, $b \in \mathbb{R}^n$, and the entries $A_{i,j} \sim \mathcal{N}(0, 1)$ are independent and identically distributed. Explain why:

1. for $d \geq 2$, bad conditioning is possible, even though the entries of A are Gaussian
2. almost surely, f satisfies the Polyak–Łojasiewicz inequality $\|\nabla f(x)\|^2 \geq 2\mu_+(f(x) - f^*)$, where μ_+ is the smallest positive eigenvalue of $A^\top A$

Exercise 2 Prove corollary 1.1.

Answer of exercise 2

Proof by contradiction. Suppose that $x^{(t)}$ does not converge towards x^* , i.e.,

$$\exists \varepsilon > 0, \forall T \in \mathbb{N}, \exists t \geq T, \quad \|x^{(t)} - x^*\| > \varepsilon.$$

Using boundedness from step (ii), extract a convergent subsequence staying at distance at least ε from x^* . By step (iii), its limit is a critical point distinct from x^* . Conclude that you have two critical points.

Exercise 3 Show that the coefficient $2\alpha(1 - \alpha)/L$ in (1.5) is maximized for $\alpha = \frac{1}{2}$.

References

- Bellman, Richard E. (1961). *Adaptive Control Processes*. Princeton University Press.
- Cauchy, Augustin (1847). “Méthode Générale Pour La Résolution Des Systèmes d’équations Simultanées”. In: *C.R. Acad. Sci. Paris* 25, pp. 536–538.
- Donoho, David L (2000). “High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality”. AMS Math Challenges Lecture.
- Nemirovski, Arkadi and David Yudin (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley. 388 pp.
- Nesterov, Yurii (2018). *Lectures on Convex Optimization*. Vol. 137. Springer.

A. Propositions used from Lecture 1

We recall here the propositions from Lecture 1 used in this lecture. All norms in this appendix are Euclidean.

Proposition A.1: Differential characterization of convexity

Let f be a differentiable function on an open set $\Omega \subseteq \mathbb{R}^p$, and $C \subseteq \Omega$ a convex subset of Ω . Then f is convex on C if, and only if,

$$\forall x, \bar{x} \in C, \quad f(x) \geq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle.$$

Proposition A.2: Quadratic upper-bound

Let f a L -smooth function, with $L > 0$. Then, for all $x, y \in \text{dom} f$,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L \|x - y\|^2. \quad (\text{A.1})$$

Moreover, if $\text{dom} f$ is convex, then (A.1) is equivalent to

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

Proposition A.3: Co-coercivity of the gradient

Let f be a full-domain convex L -smooth function, with $L > 0$. Then, ∇f is $1/L$ -co-coercive, i.e.,

$$\forall x, y \in \mathbb{R}^n, \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2.$$

Proposition A.4: Quadratic lower-bound at a minimizer

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a differentiable and μ -strongly convex function, with $\mu > 0$. Then,

$$\forall x, \bar{x} \in \mathbb{R}^n, \quad f(x) \geq f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{\mu}{2} \|x - \bar{x}\|^2.$$

In particular, if x^* is the minimizer of f , then

$$f(x) \geq f(x^*) + \frac{\mu}{2} \|x - x^*\|^2. \quad (\text{A.2})$$

Proposition A.5: Strengthened co-coercivity

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ have full effective domain and be L -smooth and μ -strongly convex, where $0 < \mu \leq L$. For any $x, y \in \mathbb{R}^n$, we have

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{\mu + L} \|x - y\|^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|^2.$$