Following the steepest direction

Samuel Vaiter

2024-02-08

Contents

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 $\hat{x} \in \mathbb{R}^d$ of $x^\star = 0$ such that $\|\hat{x} - x^\star\| \leq \epsilon$. 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

$$G_\epsilon = \{k\epsilon \mid k \in \{-\lfloor \epsilon^{-1} \rfloor, \dots, \lfloor \epsilon^{-1} \rfloor\}\},$$

and $G_\epsilon^d = G_\epsilon \times \cdots \times G_\epsilon$ its counterpart in dimension d. It is easy to see that taking $\hat{x}$ the minimizers of f over G_ϵ^d satisfies the property $\|\hat{x} - x^\star\| \leq L\epsilon$ as soon as f is L-Lipchitz. So what's the deal?

The issue is that the size of G_ϵ^d grows exponentially fast with the dimension d. In dimension 1, one needs to perform $2\lfloor \epsilon^{-1} \rfloor$ evaluations of f, but in dimension d, one needs $2\lfloor \epsilon^{-1} \rfloor^d$! This quantity gets prohibilitively large in a too fast manner: say you look at a very modest precision of $\epsilon = 10^{-2}$, then in dimension $d = 1$, only 200 computations of f are needed, whereas in dimension $d = 10$, you already need 20000000000000000000 evaluations (that's 20 zeros).

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

1 Descent direction

Consider the class of algorithms of the form

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

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

Definition 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$. A **descent direction** $u \in \mathbb{R}^d$ of f at $x \in \mathbb{R}^d$ is vector such that*

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

Observe that if f is differentiable, this property is equivalent to $\langle \nabla f(x), u \rangle < 0$. The negative gradient $-\nabla f(\bar{x})$ is the *direction of steepest descent* at the point $\bar{x}$.

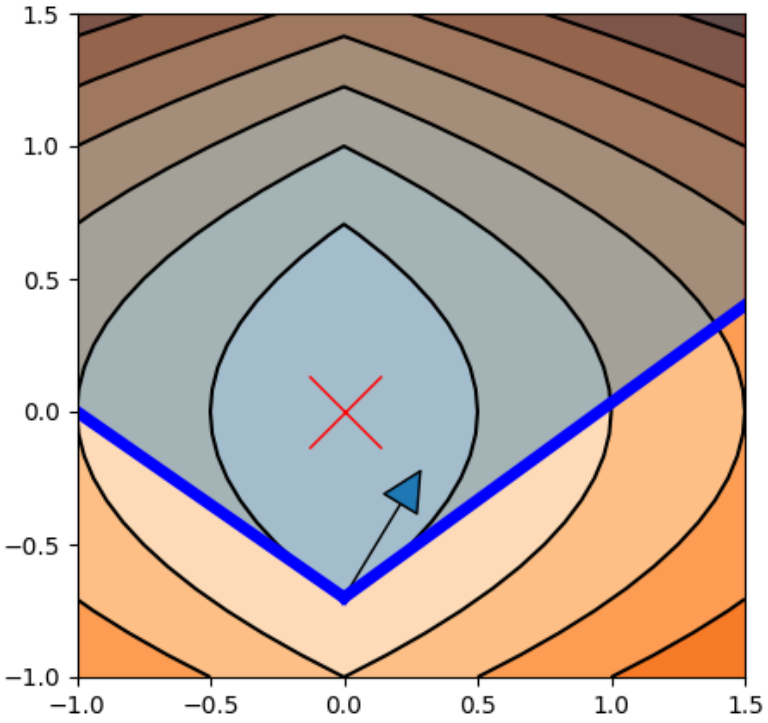

Figure 1: Descent cone.

Proposition 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function at $\bar{x} \in \mathbb{R}^d$. Then, the problem*

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

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

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

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

The function ϕ 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

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

□

2 The gradient descent (GD) algorithm

Proposition 1 gives birth to what is known as the *gradient descent algorithm* or *gradient descent method*.

Require: Initialization $x^{(0)} \in \mathbb{R}^d$, step-size policy $\eta^{(t)} > 0$.
  for $t = 1, \cdots$ **do**
    $x^{(t+1)} \leftarrow x^{(t)} - \eta^{(t)} \nabla f(x^{(t)})$
  end for

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

- *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., $g(t) = \eta(k+1)^{-1/2}$ for some $\eta > 0$. For some classes of optimization problem, it is possible to provide guarantees depending on the choice of g.

- *Oracle.* This is the "optimal" descent that a gradient descent method can produce, defined by

$$\eta_{\text{oracle}}^{(t)} = \arg\min_{\eta \geq 0} f(x^{(t)} - \eta \nabla f(x^{(t)})).$$

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

- *Backtracking rule.* Another time...

3 Rates of convegence

3.1 Means of convergence

In optimization, we consider several types of (worst-case) convergences:

- *Convergence in norms of the iterates.*

Assuming that there exists a minimizer $x^\star$, the convergence in norms (for a given norm, typically the Euclidean one) looks at an ϵ-solution in the sense of

$$r^{(t)} = \|x^{(t)} - x^\star\| \leq \epsilon.$$

Note that there may be several minimizers, but we only consider **one** given minimzier.

- *Convergence to the set of minimizers.*

Here, we look at the distance to *any* minimizer:

$$r^{(t)} = d(x^{(t)}, \mathcal{S}) = \inf_{x^\star \in \mathcal{S}} \|x^{(t)} - x^\star\| \leq \epsilon,$$

where $\mathcal{S}$ is the set of minimizer.

- *Convergence in value (0-order).*

Assuming that the infimum of the problem is given by $f^* = \inf_{x \in \text{dom} f} f(x)$, an ϵ-solution is given by

$$r^{(t)} = f(x^{(t)}) - f^* \leq \epsilon,$$

and assuming $x^\star$ is a solution, this is the quantity $r^{(t)} = f(x^{(t)}) - f(x^\star)$.

- *First-order convergence.*

If f is differentiable, one can looks at the gradient of the iterates as

$$r^{(t)} = \|\nabla f(x^{(t)})\| \leq \epsilon.$$

The idea is to use the first-order condition, and say that the more $\nabla f(x^{(t)})$, the closer we are from a critical point of f if f is nonconvex, and closer from a global minima if f is convex.

3.2 Rates versus complexity

To be more precise, we consider R-rate below, there is an alternative definition (Q-convergence).

- *Sublinear rates.*

We say that $r^{(t)}$ enjoys a *sublinear rate* if it has an upper bound written as a power function of t, i.e. $r^{(t)} \leq ct^{-\alpha}$ where $c > 0$ and $\alpha > 0$. Typical examples are stochastic gradient descent enjoying a $O(1/\sqrt{t})$ rate for convex functions or the gradient descent enjoying a $O(1/t)$ rate for convex L-smooth functions. In term of *complexity*, if $r^{(t)}$ enjoys a $ct^{-\alpha}$ rate, its complexity bound is $(\frac{c}{\epsilon})^{1/\alpha}$. For instance, the SGD for convex functions has complexity $O((\frac{c}{\epsilon})^2)$.

- *Linear rates.*

We say that $r^{(t)}$ enjoys a *linear rate* if it has an upper bound written as an exponential function of t, i.e. $r^{(t)} \leq ce^{-qt}$ for $c > 0$ and $q \in (0, 1]$. Note that typically, we don't obtain directly the exponential function in proofs, but an even more precise one: $r^{(t)} \leq c(1-q)^t$ that is directly bounded by ce^{-qt}. In term of complexity, we have the bound $\frac{1}{q}(\log c + \log \frac{1}{\epsilon})$.

- *Quadratic rates.*

We say that $r^{(t)}$ enjoys a *quadratic rate* if it has an upper bound written as

$$r^{(t+1)} \leq c(r^{(t)})^2$$

for $c > 0$. In terms of complexity, we have the bound $\log \log \frac{1}{\epsilon}$: each iteration doubles the precision of the estimate! Note that quadratic rates are a special case of a wider class known as *superlinear* convergence.

4 Explicit analysis of the gradient descent: ordinary least-squares

This part is heavily inspired from (Bach, Francis, 2021, chapter V). 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

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

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). The first-order condition for a potential minimizer $x^\star$ is given by

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

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

$$x^{(t)} - x^\star = (I - \eta H)^t (x^{(0)} - x^\star). \tag{2}$$

Proof. Using the expression of the gradient, we have for any minimizer $x^\star$

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

Substracting $x^\star$ from both side gives

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

This identity is linear recursion, and unrolling it leads to the result. □

4.1 Convergence in norm

We denote by μ the smallest eigenvalue of H, L its largest, and we let $\kappa = \frac{L}{\mu}$ its condition number. We have $0 \leq \mu \leq L$ and $\kappa \geq 1$.

Using (2), we have

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

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 such that $(I-\eta H)v = v$. Choose $x^\star$ a minimizer. Using $x^{(0)} = v + x^\star$, we have

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

Hence, the method is not necessarly convergent, depending on the initialization.

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

Proposition 2. *Assume that* (1) *has a unique minimizer* $x^\star$*. Then, the gradient descent iterates* $x^{(t)}$ *with constant step-size* $\eta^{(t)} = \eta = \frac{2}{\mu+L}$ *converges linearly in norm*

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

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

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

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,

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

Hence,

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

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 stepsize (hence slower) by taking the suboptimal 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

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

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

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

Thus, to obtain a fraction $\epsilon\|x^{(0)} - x^\star\|^2$ it is necessary to perform t iterations defined by

$$\epsilon = \exp(-\frac{2t}{\kappa}) \Leftrightarrow t = \frac{\kappa}{2}\log\frac{1}{\epsilon}.$$

4.2 Convergence in value

As we say, if $\mu = 0$, one cannot establish the convergence of the iterates. But the story is different for the convergence in value.

Proposition 3. *Let $x^\star$ any solution of (1). The gradient descent iterates $x^{(t)}$ with constant step-size $\eta^{(t)} = \eta = \frac{1}{L}$ converges with a $O(1/t)$ rate*

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

If moreover, (1) has a unique solution, the convergence is linear:

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

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

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

Using the (exact) Taylor expansion (3) of f, we have

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

Using (2), we have

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

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

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

Using the fact that for a symmetric matrix, we have $\langle Au, v\rangle \leq \|A\|_{sp}\langle u, v\rangle$, we have

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

Using again (3), we have

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

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

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

Case $\mu = 0$. Let's study the eigenvalues of $[^{\text{gradient}}_{\text{descent}}\text{-}1]$ $(I - \eta H)^{2t} H$.

We keep our analysis restricted to $\eta \leq \frac{1}{L}$.

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

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

[gradient descent-1]: Note the presence of H *after* $(I - \eta H)^{2t}$.

5 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. We start with some "weak" results on function with limited regularity and without convexity assumptions.

5.1 Coercive continuously differentiable function

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

Proposition 4. Convergence of GD for coercive C^1 functions

*Let $f : \mathbb{R}^d \to \mathbb{R}$ 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. **Step 1. 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

$$\eta^{(t)} = \arg\min_{\eta \geq 0} \phi^{(t)}(\eta) := f(x^{(t)} - \eta \nabla f(x^{(t)})).$$

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 first-order condition reads $\frac{d}{d\eta}\phi^{(t)}(0) = 0$. 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 $-\langle \nabla f(x^{(t)}), \nabla f(x^{(t)})\rangle = 0$, hence $\nabla f(x^{(t)}) = 0$ that means $x^{(t)}$ is a critical point.

Step 2. The iterates sequence is converging (up to a subsequence). Consider the set $S = \{x \in \mathbb{R}^d \mid f(x) \leq f(x^{(0)})\}$. Remark that:

- S is defined by an inequality $\leq$, 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 $\psi : \mathbb{N} \to \mathbb{N}$ is increasing.

Step 3. The limit of the iterates sequence is a critical point. Suppose that x^* is not a critical point. Using **Step 1**, 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 contiously differentiable function f, we have

$$\lim_{k\to+\infty} \left(y^{(t)} - \eta^* \nabla f(y^{(t)})\right) = x^* - \eta^* \nabla f(x^*) = x^{**}. \tag{3}$$

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

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

Hence,

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

Using (3), we get that $f(x^{**}) \geq f(x^\star)$, that is a contradiction. In conclusion, x^* is a critical point. □

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

Corollary 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a coercive, continuously differentiable function with a unique global minimizer $x^\star$. For any initial guess $x^{(0)}$, the* oracle *GD iterates $x^{(t)}$ converge towards $x^\star$*

5.2 Reminders on L-smooth functions

Recall that a (full-domain) differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be L-smooth if ∇f is L-Lipschitz. L-smoothness is important in optimization because it ensures a quadratic upper bound of the function.

Proposition 5. *Let f a L-smooth function. Then, for all $x, y \in \mathbb{R}^n$,*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L\|x - y\|^2. \tag{4}$$

Moreover, (4) *is equivalent to*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n. \tag{5}$$

5.3 Smooth function bounded from below

The previous result was rather weak. Consider now a function $f : \mathbb{R}^d \to \mathbb{R}$ that is L-smooth, but not necessarly convex. In addition, we suppose that f has full domain and is bounded from 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 6. Convergence of GD for L-smooth functions

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a L-smooth function, bounded from below by $\bar{f} \in \mathbb{R}$. For any initial guess $x^{(0)}$, the GD iterates $x^{(t)}$ with step-size $\eta^{(t)} = \eta < \frac{2}{L}$ converge towards a critical point of f, 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 $\omega = 2\alpha(1-\alpha)$ and $\eta = \frac{2\alpha}{L}$.

Proof. From Nesterov p.29

Step 1. Bound on one step of gradient descent. For $x \in \mathbb{R}^d$, let $y = x - \eta \nabla f(x)$. Using Proposition 5, we have that

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

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

$$f(x) - f(y) \geq \frac{2\alpha(1-\alpha)}{L}\|\nabla f(x)\|^2. \tag{6}$$

Step 2. Convergence of 1st-order optimality. Applying (6) to $x = x^{(t)}$ and $y = x^{(t+1)}$, we get that

$$f(x^{(t)}) - f(x^{(t+1)}) \geq \frac{2\alpha(1-\alpha)}{L}\|\nabla f(x^{(t)})\|^2. \tag{7}$$

Summing inequality (7) 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 r.h.s. telescopes, we have

$$\frac{2\alpha(1-\alpha)}{L} \sum_{t=0}^{T} \|\nabla f(x^{(t)})\|^2 \leq f(x^{(0)}) - f(x^{(T+1)}) \leq f(x^{(0)}) - \bar{f}, \tag{8}$$

by definition of $\bar{f}$. Since (8) is true for all $T \in \mathbb{N}$, we get $\nabla f(x^{(t)})$ converges towards 0.

Step 3. Rate of convergence. Using the fact that for all $t = 0 \dots T$, we have $\|\nabla f(x^{(t)})\| \leq \min_{0 \leq t \leq T} \|\nabla f(x^{(t)})\|$, we have from (8) 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. □

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 rate with respect to the bound in (6).

6 Convergence of gradient descent for convex functions

6.1 Co-coercivity of the gradient of a convex L-smooth function

Convex L-smooth functions are co-coercive. This result is sometimes coined Baillon–Haddad theorem [@baillon1977quelquesproprietes], but one shall note that the original contributionis much more general than its application to the gradient.

Proposition 7. *Let f be a full-domain convex L-smooth function. Then, ∇f is $1/L$-co-coercive, i.e.,*

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

6.2 Convex L-smooth function

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

Proposition 8. Rate of GD for convex L-smooth functions

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex L-smooth function. 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 $x^\star$ of f, and moreover

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

Moreover, if $\eta = \frac{1}{L}$, then

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

Proof. From Nesterov p.80

Let $x^\star$ a minimizer of f. We consider $r^{(t)} = \|x^{(t)} - x^\star\|$ the lack of optimality [1]. We have

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

[1] Note that we do not assume that $x^{(t)}$ converges towards this specific $x^\star$.

Using the co-coercivity of the gradient (Proposition 7), we have that

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

Thus,

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

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

Using the smoothness of f, Proposition 5 gives us the bound

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

Using again the co-coercivity of the gradient, we have that

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

Now, let $\delta^{(t)} = f(x^{(t+1)}) - f(x^\star)$. Using the differential characterization of convexity

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

, we have

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

Using Cauchy–Schwarz inequality, we have

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

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 $\frac{\delta^{(t)}}{\delta^{(t+1)}} \leq 1$, hence

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

Summing $T-1$ 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 5, we have

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

Thus,

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

which is the claimed result. □

6.3 Strongly convex L-smooth function

When f is both L-smooth and μ-strongly convex, we have the following "strenghened" co-coercivity.

Proposition 9. *Let f be a L-smooth and μ-strongly convex function. For any $x, y \in \mathbb{R}^n$, we have*

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

We now prove the **linear** convergence of gradient descent when dealing with strongly convex functions.

Proposition 10. Rate of GD for strongly convex smooth functions

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

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

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

$$\|x^{(t)} - x^\star\|^2 \leq \left(\frac{K_f - 1}{K_f + 1}\right)^{2t} \|x^{(0)} - x^\star\|^2$$
$$f(x^{(t)}) - f^\star \leq \frac{L}{2}\left(\frac{K_f - 1}{K_f + 1}\right)^{2t} \|x^{(0)} - x^\star\|^2,$$

where $K_f = L/\mu$.

Proof. From nesterov p.101 Let $r^{(t)} = \|x^{(t)} - x^\star\|$. We start like the proof of Proposition 8:

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

Using Proposition 9, we have

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

that is

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

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 $\eta - \frac{2}{\mu+L} < 0$, 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 5. □

Bach, Francis (2021). *Learning Theory from First Principles.*
Bellman, Richard E. (1961). *Adaptive Control Processes*, Princeton University Press.
Donoho, David L (2000). *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality*, Citeseer.