On gradient descent

Irène Waldspurger*

October 2023

In the whole lecture, we imagine that we want to find a minimizer of a function $f : \mathbb{R}^n \to \mathbb{R}$:

$$\text{find } x_* \text{ such that } f(x_*) = \min_{x \in \mathbb{R}^n} f(x). \tag{1}$$

We assume that at least one minimizer exists (which is for example guaranteed if f is continuous and coercive[1]) and denote one of them by x_*.

Throughout the lecture, we will assume that f is differentiable. Minimizing non-differentiable functions is is called *non-smooth optimization*. It is of course also of interest, but it requires a specific theory, which we will not have time to cover here.

1 Basic gradient descent

1.1 Reminders

Definition 1.1

For any x, the gradient of f at x is

$$\nabla f(x) \overset{def}{=} \left(\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x) \right) \in \mathbb{R}^n.$$

(It exists, because we have assumed that f is differentiable.)

*waldspurger@ceremade.dauphine.fr

[1] or even if f is only lower-semicontinuous and coercive

If f is twice differentiable, we also define its Hessian at any point x as

$$\text{Hess}\, f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n}.$$

As explained in a previous lecture, the gradient at a point $x \in \mathbb{R}^n$ provides a linear approximation of f in a neighborhood of f: informally,

$$\forall y \text{ close to } x, \qquad f(y) \approx f(x) + \langle \nabla f(x), y - x \rangle \,. \tag{2}$$

Consequently, $-\nabla f(x)$ is the direction along which f decays the most around x. This motivates the definition of gradient descent: starting at any $x_0 \in \mathbb{R}^n$, we define $(x_t)_{t \in \mathbb{N}}$ by

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t), \qquad \forall t \in \mathbb{N}.$$

Here α_t is a positive number, called the *stepsize*. In this lecture, we will restrict ourselves to constant stepsizes, except in Subsection 1.5, where we discuss better ways to choose the stepsize.

Input: A starting point x_0, a number of iterations T, a sequence of stepsizes $(\alpha_t)_{0 \leq t \leq T-1}$
for $t = 0, \ldots, T-1$ **do**
  Define $x_{t+1} = x_t - \alpha_t \nabla f(x_t)$.
end
Output: x_T

Algorithm 1: Gradient descent

Since our goal is to find a minimizer of f, we hope that

$$x_t \overset{t \to +\infty}{\to} x_*$$

or, at least,

$$f(x_t) \overset{t \to +\infty}{\to} f(x_*)$$

The goal of today's lecture in to understand under which assumptions on f we can guarantee that this happens, and, when it does, what is the convergence rate.

Before stating the main results, let us review what you have seen in the previous lectures about the convergence of gradient descent when f is quadratic.

Let $n > 0$ be an integer, C a symmetric $n \times n$ matrix, and $b \in \mathbb{R}^n$ a vector. Let f be defined as

$$\forall x \in \mathbb{R}^n, \quad f(x) = \frac{1}{2} \langle x, Cx \rangle + \langle x, b \rangle .$$

We assume that f is convex, which is equivalent to

$$C \succeq 0.$$

In this case, you have seen that, when $\lambda_{\min}(C) > 0$, gradient descent converges to a minimizer and the convergence rate is geometric (that is, fast). When $\lambda_{\min}(C) = 0$, this may not be true but $(f(x_t))_{t \in \mathbb{N}}$ nevertheless converges to $(f(x_*))$, with convergence rate at least $O(1/t)$. This is what the following theorem says.

Theorem 1.2

Let us consider the sequence of iterates $(x_t)_{t \in \mathbb{N}}$ generated by gradient descent with constant stepsize $\alpha < \frac{2}{\lambda_{\max}(C)}$.

- If $\lambda_{\min}(C) > 0$, it holds for any t that

  $$f(x_t) - f(x_*) \leq \rho^t (f(x_0) - f(x_*))$$

  for some $\rho \in]0; 1[$.

  (Actually, you have even seen that the sequence of iterates $(x_t)_{t \in \mathbb{N}}$ converges geometrically to x_*.)

- Even if $\lambda_{\min}(C) = 0$, it holds for any t that

  $$f(x_t) - f(x_*) \leq \frac{||x_0 - x_*||}{4\tau t}.$$

1.2 Convergence guarantees for general functions

The goal of this lecture is to extend to general convex functions the results stated in the quadratic case. More precisely, we will show the following guarantees.

- When f is convex and ∇f is Lipschitz, $(f(x_t))_{t\in\mathbb{N}}$ goes to $f(x_*)$ at speed $O\left(\frac{1}{t}\right)$ (Theorem 1.11). This result generalizes the situation where f is quadratic and $\lambda_{\min}(C)$ may be zero.
- When f is strongly convex and ∇f is Lipschitz, $(f(x_t))_{t\in\mathbb{N}}$ goes to $f(x_*)$ at a geometric rate (Theorem 1.14). This result generalizes the situation where f is quadratic and $\lambda_{\min}(C) > 0$.

1.2.1 Smooth functions

Let us first see what we can say of the behavior of gradient descent without assuming that f is convex. Consequently, we let f be a general differentiable function, and make only one hypothesis: f has some amount of regularity. More precisely, we assume that ∇f is Lipschitz.

Definition 1.3 : smoothness

For any $L > 0$, we say that f is L-smooth if ∇f is L-Lipschitz, that is

$$\forall x, y \in \mathbb{R}^n, \quad ||\nabla f(x) - \nabla f(y)|| \leq L||x - y||.$$

Remark

For any $L > 0$, when f is twice differentiable, it is L-smooth if and only if, for any $x \in \mathbb{R}^n$,

$$|||\text{Hess}\, f(x)||| \leq L.$$

[The notation $|||.|||$ stands for the operator norm: for any symmetric $n \times n$ matrix C, $|||C||| = \sup_{||u||_2=1} ||Cu||_2 = \max\left(|\lambda_{\min}(C)|, |\lambda_{\max}(C)|\right)$.]

Proof. Let us assume f to be twice differentiable.

If f is L-smooth, then, for any $x \in \mathbb{R}^n$, it holds for any $h \in \mathbb{R}^n$ that

$$\begin{aligned} |\left\langle \text{Hess}\, f(x)h, h\right\rangle| &= \left|\lim_{\epsilon\to 0} \frac{1}{\epsilon}\left\langle \nabla f(x+\epsilon h) - \nabla f(x), h\right\rangle\right| \\ &\leq ||h|| \limsup_{\epsilon\to 0} \frac{||\nabla f(x+\epsilon h) - \nabla f(x)||}{\epsilon} \\ &\leq L||h||^2, \end{aligned}$$

which implies that $|||\text{Hess}\, f(x)||| \leq L$.

Conversely, if $|||\text{Hess}\, f(x)||| \leq L$ for any $x \in \mathbb{R}^n$, it holds for any $x, y \in \mathbb{R}^n$ that

$$\begin{aligned} ||\nabla f(x) - \nabla f(y)|| &= \left|\left| \int_0^1 \text{Hess}\, f(x + t(y-x))(y-x)dt \right|\right| \\ &\leq \int_0^1 |||\text{Hess}\, f(x + t(y-x))|||\, ||y-x||dt \\ &\leq L||x-y|| \int_0^1 1dt \\ &= L||x-y||. \end{aligned}$$

□

Example 1.4

For any L, our quadratic function $f : x \to \frac{1}{2}\langle x, Cx\rangle + \langle x, b\rangle$ is L-smooth if and only if

$$|||C||| \leq L,$$

that is $-L \leq \lambda_{\min}(C) \leq \lambda_{\max}(C) \leq L$.

When f is smooth, the main two statements about gradient descent (with suitable constant stepsize) are given by Corollary 1.7.

- $(f(x_t))_{t\in\mathbb{N}}$ is nonincreasing (in particular, it converges);
- $(\nabla f(x_t))_{t\in\mathbb{N}}$ goes to 0.

Let us state and prove these results.

Lemma 1.5

Let $L > 0$ be fixed. If f is L-smooth, then, for any $x, y \in \mathbb{R}^n$,

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

Proof. For any $x, y \in \mathbb{R}^n$,

$$\begin{aligned}
f(y) &= f(x) + \int_0^1 \langle \nabla f(x + t(y-x)), y - x \rangle \, dt \\
&= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), y - x \rangle \, dt \\
&\leq f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 ||\nabla f(x + t(y-x)) - \nabla f(x)|| \, ||y - x|| dt \\
&\leq f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 Lt||y - x||^2 dt \\
&= f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}||y - x||^2.
\end{aligned}$$

□

Corollary 1.6

Let f be L-smooth, for some $L > 0$.
We consider gradient descent with constant stepsize: $\alpha_t = \frac{1}{L}$ for all t.
Then, for any t,

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

Corollary 1.7

With the same hypotheses as in the previous corollary, and additionally assuming that f is lower bounded,

1. $(f(x_t))_{t \in \mathbb{N}}$ converges to a finite value;
2. $||\nabla f(x_t)|| \overset{t \to +\infty}{\to} 0$.

Proof. The first property holds because, from Corollary 1.6, $(f(x_t))_{t \in \mathbb{N}}$ is a non-increasing sequence, which is lower bounded because f is. The second one is because, from the same corollary,

$$\forall t \in \mathbb{N}, \quad ||\nabla f(x_t)||^2 \leq 2L \left(f(x_t) - f(x_{t+1}) \right).$$

Therefore, for any $T \in \mathbb{N}$,

$$\sum_{t=0}^{T-1} ||\nabla f(x_t)||^2 \leq 2L \left(f(x_0) - f(x_T)\right) \leq 2L(f(x_0) - \inf f).$$

Therefore, the sum $\sum_{t\geq 0} ||\nabla f(x_t)||^2$ converges, and $(||\nabla f(x_t)||)_{t\in\mathbb{N}}$ must go to zero. □

The guarantee that $||\nabla f(x_t)|| \to 0$ when $t \to +\infty$ is quite weak (although useful in some settings, as we will see in the lecture on non-convex optimization). In particular, it does not imply that $(f(x_t))_{t\in\mathbb{N}}$ converges to $f(x_*)$. To guarantee convergence to $f(x_*)$, we need stronger assumptios on f. This is where convexity comes into play.

1.3 Smooth convex functions

Definition 1.8

We say that f is convex if

$$\forall x, y \in \mathbb{R}^n, t \in [0; 1], \quad f((1-t)x + ty) \leq (1-t)f(x) + tf(y). \tag{3}$$

Proposition 1.9

When f is differentiable, it is convex if and only if

$$\forall x, y \in \mathbb{R}^n, \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle . \tag{4}$$

Convexity is a strong structural property. From Equations (3) and (4), if we have access to the value of f and ∇f at a few points, then we have upper and lower bounds for the value of f at many other points. This allows to precisely estimate the minimum and minimizer of f from only a few values. This is why optimization is possible for convex functions, while it is quite difficult for non-convex ones.

Remark

When f is twice differentiable, it is convex if and only if, for any $x \in \mathbb{R}^n$,

$$\text{Hess } f(x) \succeq 0.$$

Example 1.10

The quadratic function $f : x \to \frac{1}{2}\langle x, Cx\rangle + \langle x, b\rangle$ is convex if and only if $C \succeq 0$.

As announced, if we assume that f, in addition to being smooth, is convex, we can prove that $(f(x_t))_{t\in\mathbb{N}}$ converges to $f(x_*)$. Moreover, we have guarantees on the convergence rate, as described by the following theorem.

Theorem 1.11

Let f be convex and L-smooth, for some $L > 0$.
We consider gradient descent with constant stepsize: $\alpha_t = \frac{1}{L}$ for all t.
Then, for any $t \in \mathbb{N}$,

$$f(x_t) - f(x_*) \leq \frac{2L||x_0 - x_*||^2}{t+4}.$$

Proof. First step: We show that the sequence of iterates gets closer to the minimizer x_* at each step: For any $t \in \mathbb{N}$,[2]

$$||x_* - x_{t+1}|| \leq ||x_* - x_t||.$$

Let t be fixed. We find upper and lower bounds for $f(x_*)$ using the convexity and L-smoothness of f. First, by convexity,

$$f(x_*) \geq f(x_t) + \langle \nabla f(x_t), x_* - x_t\rangle = f(x_t) + L\langle x_t - x_{t+1}, x_* - x_t\rangle.$$

Then, using L-smoothness through Corollary 1.6, and also the fact that x_* is a minimizer of f,

$$\begin{aligned} f(x_*) &\leq f(x_{t+1}) \\ &\leq f(x_t) - \frac{1}{2L}||\nabla f(x_t)||^2 \\ &= f(x_t) - \frac{L}{2}||x_{t+1} - x_t||^2. \end{aligned}$$

[2] We do not need it for our proof, but a stronger inequality actually holds: $\forall t \in \mathbb{N}, ||x_* - x_{t+1}||^2 \leq ||x_* - x_t||^2 - ||x_{t+1} - x_t||^2$.

Combining the two bounds yields

$$\begin{aligned} f(x_t) + L\left\langle x_t - x_{t+1}, x_* - x_t\right\rangle \leq f(x_*) \leq f(x_t) - \frac{L}{2}||x_{t+1} - x_t||^2 \\ \Rightarrow \quad 2\left\langle x_t - x_{t+1}, x_* - x_t\right\rangle + ||x_{t+1} - x_t||^2 \leq 0 \\ \iff \quad ||x_* - x_{t+1}||^2 \leq ||x_* - x_t||^2. \end{aligned}$$

Second step: We can now find an inequality relating $f(x_{t+1}) - f(x_*)$ and $f(x_t) - f(x_*)$ which, applied iteratively, will prove the result. First, from corollary 1.6,

$$f(x_{t+1}) - f(x_*) \leq f(x_t) - f(x_*) - \frac{1}{2L}||\nabla f(x_t)||^2. \tag{5}$$

In addition, because f is convex, as we have already seen in the first part,

$$f(x_t) - f(x_*) \leq \left\langle \nabla f(x_t), x_t - x_*\right\rangle.$$

Using now Cauchy-Schwarz as well as the first step of the proof:

$$f(x_t) - f(x_*) \leq ||\nabla f(x_t)||\,||x_t - x_*|| \leq ||\nabla f(x_t)||\,||x_0 - x_*||.$$

In other words, $||\nabla f(x_t)|| \geq \frac{f(x_t)-f(x_*)}{||x_0-x_*||}$. We plug this into Equation (5):

$$f(x_{t+1}) - f(x_*) \leq f(x_t) - f(x_*) - \frac{1}{2L}\frac{(f(x_t) - f(x_*))^2}{||x_0 - x_*||^2}.$$

Taking the inverse (and defining, by convention, $\frac{1}{0} = +\infty$), we get

$$\begin{aligned} \frac{1}{f(x_{t+1}) - f(x_*)} &\geq \frac{1}{f(x_t) - f(x_*)} \times \frac{1}{1 - \frac{1}{2L}\frac{f(x_t)-f(x_*)}{||x_0-x_*||^2}} \\ &\geq \frac{1}{f(x_t) - f(x_*)}\left(1 + \frac{1}{2L}\frac{f(x_t) - f(x_*)}{||x_0 - x_*||^2}\right) \\ &= \frac{1}{f(x_t) - f(x_*)} + \frac{1}{2L||x_0 - x_*||^2}. \end{aligned}$$

For the second inequality, we have used the fact that $\frac{1}{1-x} \geq 1 + x$ for any $x \in [0; 1]$.

Consequently, by iteration, it holds for any $t \in \mathbb{N}$ that

$$\frac{1}{f(x_t) - f(x_*)} \geq \frac{1}{f(x_0) - f(x_*)} + \frac{t}{2L||x_0 - x_*||^2}.$$

Corollary 1.6, together with the fact that $\nabla f(x_*) = 0$, ensures that

$$f(x_0) - f(x_*) \leq \frac{L}{2}||x_0 - x_*||^2,$$

so for any $t \in \mathbb{N}$,

$$\begin{aligned}\frac{1}{f(x_t) - f(x_*)} &\geq \frac{2}{L||x_0 - x_*||^2} + \frac{t}{2L||x_0 - x_*||^2} \\ &= \frac{t+4}{2L||x_0 - x_*||^2},\end{aligned}$$

that is

$$f(x_t) - f(x_*) \leq \frac{2L||x_0 - x_*||^2}{t+4}.$$

□

If we treat $||x_0 - x_*||$ as a constant, the previous theorem guarantees that $f(x_t) - f(x_*) = O(1/t)$. Therefore, if we want to find an ϵ-approximate minimizer (that is, an x_t such that $f(x_t) - f(x_*) \leq \epsilon$), we can do so with $O(1/\epsilon)$ iterations of gradient descent. This is nice for problems where we do not need a high-precision solution, but when ϵ is very small, this is too much. Unfortunately, Theorem 1.11 is essentially optimal: There are smooth and convex functions f for which the inequality is an equality (up to minor changes in the constants).

1.4 Smooth strongly convex functions

We will now see a subclass of smooth convex functions for which gradient descent converges much faster than the slow $O(1/t)$ rate described in the last section: the class of smooth *strongly convex* functions. It generalizes the case of quadratic functions when the smallest eigenvalue is strictly positive (see Example 1.13).

Definition 1.12

Let $\mu > 0$ be fixed. If f is differentiable, we say that it is μ-strongly convex if, for any $x, y \in \mathbb{R}^n$,

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

We observe that, if f is strongly convex, then it is convex. But strong convexity is a more powerful property than convexity: If we know the value and gradient at a point x of a strongly convex function, we know a quadratic lower bound for f (which, in particular, grows to $+\infty$ away from x) instead of a simple linear lower bound as for simply convex functions.

Remark

For any $\mu > 0$, a differentiable function f is μ-strongly convex if and only if the function $f_\mu : x \to f(x) - \frac{\mu}{2}||x||^2$ is convex.

Proof. The function f_μ is convex if and only if, for any $x, y \in \mathbb{R}^n$,

$$\begin{aligned}
& f_\mu(y) \geq f_\mu(x) + \langle \nabla f_\mu(x), y - x \rangle \,; \\
\iff \quad & f(y) - \frac{\mu}{2}||y||_2^2 \geq f(x) - \frac{\mu}{2}||x||_2^2 + \langle \nabla f(x) - \mu x, y - x \rangle \,; \\
\iff \quad & f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\left(||y||_2^2 - 2\langle x, y - x \rangle - ||x||_2^2\right) \,; \\
\iff \quad & f(y) \geq f(x) + \langle \nabla f(x) - \mu x, y - x \rangle + \frac{\mu}{2}||y - x||_2^2.
\end{aligned}$$

□

Remark

As a consequence from the previous remark, as well as the one following Definition 1.8, a twice differentiable function f is μ-strongly convex if and only if, for any $x \in \mathbb{R}^n$,

$$\operatorname{Hess} f(x) - \mu \mathrm{Id} \succeq 0,$$

or, in other words, all eigenvalues of $\operatorname{Hess} f(x)$ are larger than μ.

Example 1.13

We consider again the quadratic function $f : x \in \mathbb{R}^n \to \frac{1}{2}\langle x, Cx \rangle + \langle x, b \rangle$. Its Hessian at any point is C. We denote $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ the ordered eigenvalues of C. From the previous remark, if $\lambda_n > 0$, f is λ_n-strongly convex. If $\lambda_n \leq 0$, f is not μ-strongly convex, whatever the value of $\mu > 0$.

Theorem 1.14

Let $0 < \mu < L$ be fixed. Let f be L-smooth and μ-strongly convex. We consider gradient descent with constant stepsize: $\alpha_t = \frac{1}{L}$ for all t. Then, for any $t \in \mathbb{N}$,

$$||x_t - x_*||_2 \leq \left(1 - \frac{\mu}{L}\right)^t ||x_0 - x_*||_2; \tag{6}$$
$$f(x_t) - f(x_*) \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^{2t} ||x_0 - x_*||_2^2.$$

Proof. It is enough to prove Equation (6). Indeed, if this equation holds, it implies (from Lemma 1.5 and because $\nabla f(x_*) = 0$),

$$f(x_t) \leq f(x_*) + \frac{L}{2}||x_t - x_*||_2^2. \Rightarrow \quad f(x_t) - f(x_*) \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^{2t} ||x_* - x_0||_2^2.$$

To prove Equation (6), it suffices to prove that, for any $t \in \mathbb{N}$,

$$||x_{t+1} - x_*||_2 \leq \left(1 - \frac{\mu}{L}\right) ||x_t - x_*||_2.$$

Let us fix $t \in \mathbb{N}$ and establish this inequality.

Given that $x_{t+1} = x_t - \frac{1}{L}\nabla f(x_t)$, we must simply upper bound

$$||x_{t+1} - x_*||_2 = \frac{1}{L} ||\nabla f(x_t) - L(x_t - x_*)||_2$$

with a multiple of $||x_t - x_*||_2$.

We must therefore establish an inequality involving only x_t, x_* and $\nabla f(x_t)$. For this, we first look at which inequlities we can write on these quantities. In particular, we consider the inequality defining μ-strong convexity (Definition 1.12), at $x = x_t$ or $x = x_*$: for all $y \in \mathbb{R}^n$,

$$f(y) \geq f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{\mu}{2}||y - x_t||_2^2; \tag{7a}$$
$$f(y) \geq f(x_*) + \frac{\mu}{2}||y - x_*||_2^2. \tag{7b}$$

And considering also the inequality of Lemma 1.5, we have, for all $y \in \mathbb{R}^n$,

$$f(y) \leq f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{L}{2}||y - x_t||_2^2; \tag{8a}$$
$$f(y) \leq f(x_*) + \frac{L}{2}||y - x_*||_2^2. \tag{8b}$$

In particular, for all $y \in \mathbb{R}^n$, combining (7a) and (8b), it holds that

$$f(x_*) + \frac{L}{2}||y - x_*||_2^2 - f(x_t) - \langle \nabla f(x_t), y - x_t \rangle - \frac{\mu}{2}||y - x_t||_2^2 \geq 0.$$

The minimum of this expression is reached at $y = \frac{Lx_* - \mu x_t + \nabla f(x_t)}{L - \mu}$, and its value is

$$f(x_*) - f(x_t) - \frac{||\nabla f(x_t)||_2^2}{2(L-\mu)} - \left\langle \nabla f(x_t), \frac{L(x_* - x_t)}{L - \mu} \right\rangle - \frac{L\mu}{2(L-\mu)}||x_t - x_*||_2^2 \geq 0.$$

Similarly, combining (7b) and (8a), we get for all $y \in \mathbb{R}^n$

$$f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{L}{2}||y - x_t||_2^2 - f(x_*) - \frac{\mu}{2}||y - x_*||_2^2 \geq 0.$$

The minimum of this expression is reached at $y = \frac{Lx_t - \mu x_* - \nabla f(x_t)}{L - \mu}$, and its value is

$$f(x_t) - f(x_*) - \frac{||\nabla f(x_t)||_2^2}{2(L-\mu)} + \left\langle \nabla f(x_t), \frac{\mu(x_t - x_*)}{L - \mu} \right\rangle - \frac{L\mu}{2(L-\mu)}||x_t - x_*||_2^2 \geq 0.$$

If we combine the two minima, we get

$$\begin{aligned} &(L + \mu) \langle \nabla f(x_t), x_t - x_* \rangle \geq ||\nabla f(x_t)||_2^2 + L\mu||x_t - x_*||_2^2 \\ \iff \quad &\left|\left| \nabla f(x_t) - \frac{L+\mu}{2}(x_t - x_*) \right|\right|_2 \leq \frac{L-\mu}{2}||x_t - x_*||_2. \end{aligned}$$

Together with the triangular inequality, this proves the result:

$$\begin{aligned} &\frac{1}{L}||\nabla f(x_t) - L(x_t - x_*)||_2 \\ &\quad \leq \frac{1}{L}\left|\left| \nabla f(x_t) - \frac{L+\mu}{2}(x_t - x_*) \right|\right|_2 + \frac{1}{L}\left|\left| \frac{L+\mu}{2}(x_t - x_*) - L(x_t - x_*) \right|\right|_2 \\ &\quad \leq \frac{L-\mu}{2L}||x_t - x_*||_2 + \frac{L-\mu}{2L}||x_t - x_*||_2 \\ &\quad = \left(1 - \frac{\mu}{L}\right)||x_t - x_*||_2. \end{aligned}$$

□

Hence, when f is smooth and strongly convex, $(f(x_t) - f(x_*))_{t\in\mathbb{N}}$ decays geometrically, with rate at least $\left(1 - \frac{\mu}{L}\right)^2$. An ϵ-approximate minimizer can be found in $O((\log \epsilon)/\log(1 - \mu/L))$ gradient descent iterations, much less than the $O(\epsilon)$ obtained without the strong convexity assumption.

We call $\frac{L}{\mu} \geq 1$ the *condition number* of f. The closer to 1 it is, the faster the convergence.

Remark

The rate $\left(1 - \frac{\mu}{L}\right)^2$ in the previous theorem is tight, in the sense that it is not possible to establish the same theorem for a strictly smaller convergence rate. Indeed, when applied to a μ-strongly convex and L-smooth *quadratic* function, the gradient descent iterates go to zero at this exact rate.

1.5 Choice of stepsizes

Properly choosing the stepsizes $(\alpha_t)_{t\in\mathbb{N}}$ is crucial: if they are too large, then x_{t+1} is outside the domain where the approximation (2) holds, and the algorithm may diverge. On the contrary, if they are too small, x_t needs many time steps to move away from x_0, and convergence can be slow.

What a good stepsize choice is depends on the properties of f. Let us however mention some common strategies:

1. *Fixed schedule*: the stepsizes are chosen in advance; α_t generally depends on t through a simple equation, like

$$\forall t, \quad \alpha_t = \eta, \text{ for some } \eta > 0, \qquad \text{(Constant stepsize)}$$

$$\text{or} \quad \forall t, \quad \alpha_t = \frac{1}{t+1}. \qquad \text{(Monotonically decreasing stepsize)}$$

2. *Exact line search*: for any t, choose α_t such that

$$f(x_t - \alpha_t \nabla f(x_t)) = \min_{a\in\mathbb{R}} f(x_t - a\nabla f(x_t)).$$

3. *Backtracking line search*: unless f has very particular properties, it is a priori difficult to minimize f on a line. The exact line search strategy is therefore difficult to implement. Instead, one can simply choose α_t

such that $f(x_t - \alpha_t \nabla f(x_t))$ is "sufficiently smaller than $f(x_t)$" The approximation (2) implies, for α_t small enough,

$$f(x_t - \alpha_t \nabla f(x_t)) \approx f(x_t) - \alpha_t ||\nabla f(x_t)||^2.$$

If we consider that "being sufficiently smaller than $f(x_t)$" means that the previous approximation holds, up to the introduction of a multiplicative constant, the following algorithm describes a way to find a suitable α_t.

Input: Parameters $c, \tau \in]0; 1[$, maximal stepsize value a_{max}
Define $\alpha_t = a_{max}$.
while $f(x_t - \alpha_t \nabla f(x_t)) > f(x_t) - c\alpha_t ||\nabla f(x_t)||^2$ **do**
| Set $\alpha_t = \tau \alpha_t$.
end
Output: α_t

Algorithm 2: Backtracking line search

2 References

The main references used to prepare these notes are three classical books on optimization,

- *Introductory lectures on convex optimization: a basic course*, by Y. Nesterov, Springer Science & Business Media, volume 87 (2003),
- *Convex optimization*, by S. Boyd and L. Vandenberghe, Cambridge University Press (2004),
- *Optimization for data analysis*, by S. J. Wright and B. Recht, Cambridge University Press (2022).