Gradient descent

Irène Waldspurger

October 15, 2020

1 Definition of gradient descent

Let us assume that we want to find a minimizer of a function $f : \mathbb{R}^n \to \mathbb{R}$:

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

In all the lecture, we assume that a minimizer exists, and denote it x_* [1].

1.1 Definition

We also assume that f is differentiable.

Définition 1.1. *For any x, the gradient of f at x is*

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

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,

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

1. At least, we denote one of them by x_* : The minimizer may not be unique.

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*.

Algorithm 1 Gradient descent

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

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.

1.2 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.

Algorithm 2 Backtracking line search

Require: 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 while
  return α_t

In this lecture, we will restrict ourselves to constant stepsizes.

1.3 Reminder : the quadratic case

Let f be defined as

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

where M is a symmetric $n \times n$ matrix, and b belongs to $\mathbb{R}^n$.

Proposition 1.2. *The function f is twice differentiable. For any $x \in \mathbb{R}^n$,*

$$\nabla f(x) = Mx + b;$$
$$\text{Hess}\, f(x) = M.$$

We assume that f is convex, which is equivalent to M being semidefinite positive (that is, all its eigenvalues are nonnegative). In this case, you have seen in a lecture by Gabriel Peyré that, when $\lambda_{\min}(M) > 0$, gradient descent converges to a minimizer and the convergence rate is geometric (that is, fast). When $\lambda_{\min}(M) = 0$, 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.

Théorème 1.3. *Let us consider the sequence of iterates $(x_t)_{t\in\mathbb{N}}$ generated by gradient descent with constant stepsize $\alpha < \frac{2}{\lambda_{\max}(M)}$.*

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

$$||x_t - x_*|| \leq \rho^t ||x_0 - x_*||$$

  for some $\rho \in]0; 1[$.
- *Even if $\lambda_{\min}(M) = 0$, it holds for any t that*

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

2 Convergence analysis

The goal of this section is to extend to general convex functions the results you have seen in the quadratic case, and to see which convergence guarantees it is possible to establish depending on the assumptions we can make over f.

2.1 Smooth functions

To start with, let us not assume that f is convex. We only assume that f is *smooth*, in the sense of the following definition, and see what we can say of the behavior of gradient descent.

Définition 2.1. *For any $L > 0$, we say that f is L-smooth if ∇f is L-Lipschitz, that is*

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

Remarque 2.2. *When f is twice differentiable, it is L-smooth if and only if, for any $x \in \mathbb{R}^n$,*

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

[For any symmetric matrix M, $|||M|||$ denotes the spectral norm, which is equal to $\max_{k=1,\dots,n} |\lambda_k(M)|$.]

Démonstration. 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

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

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

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

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

□

Exemple 2.3. *For any L, our quadratic function $f : x \to \frac{1}{2}\langle x, Mx\rangle + \langle x, b\rangle$ is L-smooth if and only if*

$$|||M||| \leq L,$$

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

When f is smooth, it turns out that $(f(x_t))_{t\in\mathbb{N}}$ is nonincreasing. Moreover, we can analyze the decay of $f(x_t)$ at each iteration thanks to the following lemma.

Lemme 2.4. *Let $L > 0$ be fixed. If f is L-smooth, then, for any $x, y \in \mathbb{R}^n$,*

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

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

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

□

Corollaire 2.5. *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,

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

Corollaire 2.6. *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. *$||\nabla f(x_t)|| \overset{t \to +\infty}{\to} 0$.*

Démonstration. The first property holds because, from Corollary 2.5, $(f(x_t))_{t \in \mathbb{N}}$ 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 tomorrow). In particular, it does not imply that $(f(x_t))_{t\in\mathbb{N}}$ converges to $f(x_*)$. If we want to be able to guarantee that this convergence happens, we need f to satisfy a much stronger property than smoothness. The simplest and most widely studied example of such a property is *convexity*.

2.2 Smooth convex functions

Définition 2.7. *We say that f is convex if*

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

Proposition 2.8. *When f is differentiable, it is convex if and only if*

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

Remarque 2.9. *When f is twice differentiable, it is convex if and only if, for any $x \in \mathbb{R}^n$,*

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

Exemple 2.10. *The quadratic function $f : x \to \frac{1}{2}\langle x, Mx\rangle + \langle x, b\rangle$ is convex if and only if M is semidefinite positive.*

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.

Théorème 2.11. *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 $t \in \mathbb{N}$,

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

Démonstration. 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]

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

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

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 2.5, and also the fact that x_* is a minimizer of f,

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

Combining the two bounds yields

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

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 2.5,

$$f(x_{t+1}) - f(x_*) \leq f(x_t) - f(x_*) - \frac{1}{2L}||\nabla f(x_t)||^2. \tag{4}$$

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

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

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 (4) :

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

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 2.5, 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 2.11 is essentially optimal : There are smooth and convex functions f for which the inequality is an equality (up to minor changes in the constants).

2.3 Smooth strongly convex functions

In the previous paragraph, we have seen that gradient descent allows to approximately minimize any smooth convex function, but at a relatively slow rate. We will now see a subclass of smooth convex functions for which gradient descent converges much faster : the class of smooth *strongly convex* functions.

Définition 2.12. *Let $\mu > 0$ be fixed. If f is differentiable, we say that it is μ-strongly convex if, for any $x, y \in \mathbb{R}^n$,*

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

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.

Remarque 2.13. *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.*

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

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

□

Remarque 2.14. *As a consequence from Remarks 2.9 and 2.13, 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.*

Exemple 2.15. *We consider again the quadratic function* $f : x \in \mathbb{R}^n \to \frac{1}{2}\langle x, Mx\rangle + \langle x, b\rangle$. *Its Hessian at any point is* M. *We denote* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ *the ordered eigenvalues of* M. *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* $\mu > 0$.

Théorème 2.16. *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 $t \in \mathbb{N}$,

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

Démonstration. The first part of the proof is similar to the one of Theorem 2.11. In the proof of Theorem 2.11, we had shown that $(||x_t - x_*||)_{t\in\mathbb{N}}$ was a non-increasing sequence. With the same reasoning but using strong convexity instead of plain convexity, we improve this result and show that $(||x_t - x_*||)_{t\in\mathbb{N}}$ actually goes to zero at a geometric rate.

Let t be fixed. By strong convexity,

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

And using L-smoothness as in the proof of Theorem 2.11,

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

We combine the two bounds :

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

We can conclude : From Lemma 2.4 and because $\nabla f(x_*) = 0$,

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

As a consequence,

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

□

Hence, when f is smooth and strongly convex, $(f(x_t) - f(x_*))_{t\in\mathbb{N}}$ decays geometrically, with rate at least $1 - \frac{\mu}{L}$. 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.

Remarque 2.17. *The rate $1 - \frac{\mu}{L}$ in the previous theorem is not optimal. With a more sophisticated proof, we could have shown that, for any $t \in \mathbb{N}$,*

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

3 Acceleration

To conclude this lecture, we go back to the setting where f is L-smooth, for some $L > 0$, and convex. We have seen in Theorem 2.11 that $f(x_t) - f(x_*) = O(1/t)$. As we said, this theorem cannot be significantly improved without additional assumptions on f, like strong convexity : In the worst situations, gradient descent really converges at rate $O(1/t)$.

However, gradient descent may not be the best possible algorithm. Are there other algorithms, that, from only the knowledge of ∇f at some points, achieve a faster convergence rate ? The answer is yes. An example of such an algorithm has been provided by Yurii Nesterov.

Two essential ideas for understanding the algorithm are :

1. At each time step, gradient descent evaluates the gradient of f at the current iterate x_t and defines x_{t+1} from this information only. It completely discards the information obtained at previous time steps. A better method must take this previous information into account.
2. Computing the gradient of f precisely at x_t is the most intuitive choice, but maybe not the most intelligent one. There may be another point where ∇f carries more information on x_* and $f(x_*)$.

Therefore, in Nesterov's algorithm, two sequences $(x_t)_{t\in\mathbb{N}}$ and $(y_t)_{t\in\mathbb{N}}$ are defined. The first one, $(x_t)_{t\in\mathbb{N}}$, is the sequence of approximate minimizers : If we stop our algorithm at time t, it returns x_t. The second one, $(y_t)_{t\in\mathbb{N}}$, is the sequence of points at which we compute ∇f. These sequences are defined by

the following iteration formulas :

$$x_{t+1} = y_t - \frac{1}{L}\nabla f(y_t);$$
$$y_{t+1} = x_{t+1} + \gamma_t(x_{t+1} - x_t),$$

with $x_0 = y_0$ an arbitrary starting point, and where $(\gamma_t)_{t\in\mathbb{N}}$ is a carefully chosen sequence of real numbers, whose exact (and admittedly mysterious, at first sight) definition, is

$$\lambda_{-1} = 0,$$
$$\forall t \in \mathbb{N}, \quad \lambda_t = \frac{1 + \sqrt{1 + 4\lambda_{t-1}^2}}{2},$$
$$\forall t, \quad \gamma_t = \frac{\lambda_t - 1}{\lambda_{t+1}}.$$

The following theorem provides a convergence rate for Nesterov's algorithm.

Théorème 3.1. *For any $t \in \mathbb{N}$,*

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

The convergence rate of Nesterov's algorithm is therefore $O(1/t^2)$, compared to $O(1/t)$ for gradient descent. One can show that this convergence rate is optimal among all algorithms that only exploit gradient information about f (called *first-order algorithms*).

4 References

The main sources used to prepare this lecture are two classical books :

- Convex optimization, by S. Boyd and L. Vandenberghe, which is a relatively easy-to-read introduction to optimization ;
- Introductory lectures on convex optimization : a basic course, by Y. Nesterov, which is more technical and theoretical than the previous one.

For the part on acceleration, two blog posts by S. Bubek have also been useful :

- http://blogs.princeton.edu/imabandit/2013/04/01/acceleratedgradientdescent/
- http://blogs.princeton.edu/imabandit/2018/11/21/a-short-proof-for-nesterovs-momentum/