Gradient descent

Irène Waldspurger

October 8, 2019

1 Definition of gradient descent

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

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

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

1.1 Motivation and definition

An intuitively reasonable strategy to solve Problem (1) is to start from an arbitrary point $x_0 \in \mathbb{R}^n$, gather some information on f around x_0, and use it to find another point x_1, hopefully closer to a minimizer than x_0. Doing that repeatedly yields a sequence of points $(x_t)_{t\in\mathbb{N}}$. If everything goes well,

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

that is, for large t, x_t is an approximate minimizer of f.

We now assume that f is differentiable.

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

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

Proposition 1.2 (Informal). *For any $x \in \mathbb{R}^n$, the value of f around x can be approximated by*

$$\forall y, \qquad f(y) \approx f(x) + \langle \nabla f(x), y - x \rangle \,. \quad (2)$$

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

Therefore, around x_t, the direction along which f decays the most is $-\nabla f(x_t)$. A sensible definition for x_{t+1} from x_t is thus

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t),$$

where α_t is a positive number, called the *stepsize*, which controls the distance between x_{t+1} and x_t.

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 $x_T$
```

The main goal of today's lecture is to discuss under which hypotheses on f we can ensure that $f(x_t)$ goes to $f(x_*)$ when t goes to $+\infty$ and, under these hypotheses, what we can say about the convergence speed.

1.2 Example : quadratic function

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 1.3. *For any $x \in \mathbb{R}^n$,*

$$\nabla f(x) = Mx + b.$$

Démonstration. Let $x \in \mathbb{R}^n$ be fixed. We must compute, for any $k \in \{1, \dots, n\}$,

$$(\nabla f(x))_k = \frac{\partial f}{\partial x_k}(x) = \lim_{h\to 0} \frac{f(x + he_k) - f(x)}{h},$$

where e_k denotes the k-th vector of the canonical basis.

We observe that, for any $\Delta \in \mathbb{R}^n$,

$$\begin{aligned} f(x+\Delta) - f(x) &= \frac{1}{2}\langle x+\Delta, M(x+\Delta)\rangle + \langle x+\Delta, b\rangle - \frac{1}{2}\langle x, Mx\rangle - \langle x, b\rangle \\ &= \frac{1}{2}\langle \Delta, Mx\rangle + \frac{1}{2}\langle x, M\Delta\rangle + \langle \Delta, M\Delta\rangle + \langle \Delta, b\rangle \\ &= \langle \Delta, Mx+b\rangle + \frac{1}{2}\langle \Delta, M\Delta\rangle . \end{aligned}$$

Therefore, for any $k \in \{1, \dots, n\}$,

$$\begin{aligned} (\nabla f(x))_k &= \lim_{h\to 0} \frac{h\langle e_k, Mx+b\rangle + \frac{h^2}{2}\langle e_k, M\Delta\rangle}{h} \\ &= \lim_{h\to 0}\left(\langle e_k, Mx+b\rangle + \frac{h}{2}\langle e_k, M\Delta\rangle\right) \\ &= \langle e_k, Mx+b\rangle \\ &= (Mx+b)_k, \end{aligned}$$

so $\nabla f(x) = Mx + b$. □

Assuming M to be invertible, we see that the only point where the gradient is zero is $-M^{-1}b$. The only minimizer of f is therefore $x_* = -M^{-1}b$. For any t, the $t+1$-th gradient descent iterate is thus defined by

$$\begin{aligned} x_{t+1} &= x_t - \alpha_t(Mx_t + b) = (\mathrm{Id} - \alpha_t M)x_t - \alpha_t b, \\ \text{that is, } x_{t+1} - x_* &= (\mathrm{Id} - \alpha_t M)(x_t - x_*) + \alpha_t M x_* - \alpha_t b \\ &= (\mathrm{Id} - \alpha_t M)(x_t - x_*). \end{aligned}$$

1.3 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 $\alpha_t$
```

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

2 Convergence analysis

Recall that the goal of gradient descent is, after a sufficient number of steps, to obtain an approximate minimizer of f. Formally, we want

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

and, if possible, we want the convergence rate to be fast.

2.1 Smooth functions

A natural idea to understand the behavior of $(f(x_t))_{t\in\mathbb{N}}$ is to find an upper bound for $f(x_{t+1})$ that depends on $f(x_t)$, and apply it iteratively to upper bound $f(x_{t+1})$ using $f(x_0)$ only. The simplest hypothesis one can make on f to ensure that such an upper bound exists is *smoothness*.

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

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

Exemple 2.2. *We consider again our quadratic function $f : x \to \frac{1}{2}\langle x, Mx\rangle + \langle x, b\rangle$. For any $x, y \in \mathbb{R}^n$,*

$$\begin{aligned} ||\nabla f(x) - \nabla f(y)|| &= ||(Mx+b) - (My+b)|| \\ &= ||M(x-y)|| \\ &= |||M|||\,||x-y||, \end{aligned}$$

where $|||M|||$ denotes the operator norm of M. Standard results about symmetric matrices tell us that

$$|||M||| = \max\{|\lambda|, \lambda \textit{ eigenvalue of } M\} \overset{def}{=} \lambda_{max}(M).$$

As a consequence, the function f is $\lambda_{max}(M)$-smooth.

Lemme 2.3. *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 2.4. *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 2.5. *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 2.4, $(f(x_t))_{t\in\mathbb{N}}$ 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. □

Without additional assumptions on f, there is not much more that we can say about gradient descent. In particular, $(f(x_t))_{t\in\mathbb{N}}$ may not converge 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 2.6. *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 2.7. *When f is differentiable, it is convex if and only if*

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

Exemple 2.8. *When is our quadratic function $f : x \to \frac{1}{2}\langle x, Mx\rangle + \langle x, b\rangle$ convex ?*

We have seen while computing ∇f that, for any $x, y \in \mathbb{R}^n$ (setting $\Delta = y - x$ in our previous equation),

$$f(y) = f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\langle y - x, M(y - x)\rangle .$$

Therefore, f is convex if and only if, for any $x, y \in \mathbb{R}^n$, $\langle y - x, M(y - x)\rangle \geq 0$. This amounts to requiring that, for any $v \in \mathbb{R}^n$, $\langle v, Mv\rangle \geq 0$: f 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 speed at which convergence takes place, as described by the following theorem.

Théorème 2.9. *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]

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

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 2.4, 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 2.4,

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

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

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

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 (4) :

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

We can now establish the result by iteration over t. For $t = 0$, Corollary 2.4, together with the fact that $\nabla f(x_*) = 0$, ensures that

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

Then, for any $t \in \mathbb{N}$, if $f(x_t) - f(x_*) \leq \frac{2L||x_0-x_*||^2}{t+4}$,

$$\begin{aligned} f(x_{t+1}) - f(x_*) &\leq f(x_t) - f(x_*) - \frac{1}{2L}\frac{(f(x_t) - f(x_*))^2}{||x_0 - x_*||^2} \\ &\leq \frac{2L||x_0 - x_*||^2}{t+4} - \frac{\left(\frac{2L||x_0-x_*||^2}{t+4}\right)^2}{2L||x_0 - x_*||^2} \\ &= 2L||x_0 - x_*||^2\left(\frac{1}{t+4} - \frac{1}{(t+4)^2}\right) \\ &\leq \frac{2L||x_0 - x_*||^2}{t+5}. \end{aligned}$$

(To obtain the second inequality, we have used the facts that the map $x \to x - \frac{x^2}{2L||x_0-x_*||^2}$ is increasing over $]-\infty; L||x_0-x_*||^2]$ and that $\frac{2L||x_0-x_*||^2}{t+4} \leq L||x_0-x_*||^2$.) □

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 2.9 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 : This is the class of smooth *strongly convex* functions.

Définition 2.10. *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 2.11. *For any $\mu > 0$, f is μ-strongly convex if and only if the function $x \to f(x) - \frac{\mu}{2}||x||^2$ is convex.*

Exemple 2.12. *We consider again the quadratic function $f : x \in \mathbb{R}^n \to \frac{1}{2}\langle x, Mx \rangle + \langle x, b \rangle$. We assume that f is convex, that is $M \succeq 0$. Is it strongly convex ?*

We recall a central theorem from matrix theory : M can be diagonalized in an orthonormal basis. In other words, M can be written as

$$M = U \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{pmatrix} U^T,$$

where U belongs to $O_n(\mathbb{R})$ and $\lambda_1 \geq \dots \geq \lambda_n$ are the (ordered) eigenvalues of M.

Changing the basis from the canonical one to the one defined by the eigenvectors of M, we can assume that $M = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{pmatrix}$. For any $x, y \in \mathbb{R}^n$,

$$\begin{aligned} f(y) &= f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle y - x, M(y-x) \rangle \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \left(\lambda_1 (y_1 - x_1)^2 + \dots + \lambda_n (y_n - x_n)^2 \right) \\ &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \left(\lambda_n (y_1 - x_1)^2 + \dots + \lambda_n (y_n - x_n)^2 \right) \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \frac{\lambda_n}{2} ||y - x||^2. \end{aligned}$$

Therefore, if $\lambda_n > 0$, then f is λ_n-strongly convex.

Théorème 2.13. *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 2.9. In the proof of Theorem 2.9, 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 2.9,

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

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 2.3 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 $1 - \frac{\mu}{L}$. 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.

3 Example : quadratic function (again)

As previously, we consider the function $f : x \in \mathbb{R}^n \to \frac{1}{2}\langle x, Mx\rangle + \langle x, b\rangle$, and assume that

$$M = \begin{pmatrix} \lambda_1 & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & \lambda_n \end{pmatrix}.$$

We assume that $\lambda_1 \geq \cdots \geq \lambda_n > 0$, and recall that f is then λ_n-strongly convex and λ_1-smooth.

We consider gradient descent with constant stepsize $\alpha_t = \frac{1}{\lambda_1}, \forall t \in \mathbb{N}$. We recall from Subsection 1.2 that

$$\forall t \in \mathbb{N}, \quad x_{t+1} - x_* = \left(1 - \frac{1}{\lambda_1} M\right)(x_t - x_*),$$
$$\Rightarrow \quad \forall t \in \mathbb{N}, \quad x_t - x_* = \left(1 - \frac{1}{\lambda_1} M\right)^t (x_0 - x_*).$$

If we look at the k-th coordinate, for $k = 1, \dots, n$, this implies

$$(x_t - x_*)_k = \left(1 - \frac{\lambda_k}{\lambda_1}\right)^t (x_0 - x_*)_k. \tag{5}$$

As a consequence, for any $t \in \mathbb{N}$,

$$\begin{aligned}
f(x_t) - f(x_*) &= \frac{1}{2}\langle x_t, Mx_t\rangle + \langle x_t, b\rangle + \frac{1}{2}\left\langle M^{-1}b, b\right\rangle \\
&= \frac{1}{2}\left\langle x_t + M^{-1}b, M(x_t + M^{-1}b)\right\rangle \\
&= \frac{1}{2}\langle x_t - x_*, M(x_t - x_*)\rangle \\
&= \frac{1}{2}\sum_{k=1}^{n} \lambda_k (x_t - x_*)_k^2 \\
&= \frac{1}{2}\sum_{k=1}^{n} \lambda_k \left(1 - \frac{\lambda_k}{\lambda_1}\right)^{2t} (x_0 - x_*)_k^2 \\
&\leq \frac{1}{2}\sum_{k=1}^{n} \lambda_1 \left(1 - \frac{\lambda_n}{\lambda_1}\right)^{2t} (x_0 - x_*)_k^2 \\
&= \frac{\lambda_1}{2}\left(1 - \frac{\lambda_n}{\lambda_1}\right)^{2t} ||x_0 - x_*||^2.
\end{aligned}$$

This is essentially the convergence rate given by Theorem 2.13, with $\mu = \lambda_n$ and $L = \lambda_1$. The only difference is that the geometric rate is $\left(1 - \frac{\lambda_n}{\lambda_1}\right)^2$ instead of $\left(1 - \frac{\lambda_n}{\lambda_1}\right)$, but both numbers are of the same order, so this tells us that our analysis of the convergence rate is not far from optimal.

Equation (5) allows us to understand more precisely the behavior of the iterates. For any $k = 1, \dots, n$, $(x_{t,k})_{t \in \mathbb{N}}$ converges geometrically to $x_{*,k}$, and the rate is equal to $\left(1 - \frac{\lambda_k}{\lambda_1}\right)$. When λ_k is of the same order as λ_1, this is very fast. But if the condition number is large, that is

$$\frac{\lambda_1}{\lambda_n} \gg 1,$$

we can have $\lambda_k \ll \lambda_1$ for large k, so that the rate $\left(1 - \frac{\lambda_k}{\lambda_1}\right)$ is close to 1, and convergence is slower.

Therefore, after a few gradients steps, we typically have

$$x_{t,k} \approx x_{*,k}$$

for small values of k, and the remaining iterations are only necessary for the convergence of the last coordinates.

Intuitively, the problem here, when the condition number is large, is that the stepsize $\frac{1}{\lambda_1}$ is well-suited to the first coordinates, along which the gradient is large, but too small for the last coordinates, along which the gradient is small. This issue can be overcame with *second-order methods*, which exploit the information given by second-order derivatives and not only by the gradient, but are generally much more computationally expensive.

4 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 2.9 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 :

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

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

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

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

Théorème 4.1. *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*).

5 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/`