Non-convex optimization

Irène Waldspurger

November 26, 2019

1 Introduction

Let us consider a general unconstrained minimization problem :

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

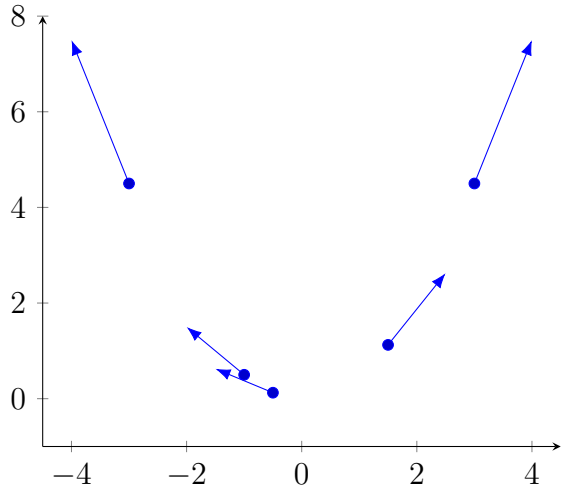
for some function $f : \mathbb{R}^n \to \mathbb{R}$. We assume that at least one minimizer exists, which we call x_*. We also assume througout the lecture that f is $\mathcal{C}^\infty$, to avoid all possible regularity issues.

In the past lectures, we have seen how to find a good approximation of a minimizer, under the assumption that f is convex. The goal of this lecture is to study what we can do when f is not convex.

1.1 Why non-convex optimization is difficult

We first try to give an intuition of the difference between convex and non-convex optimization, and why the latter one is much more difficult.

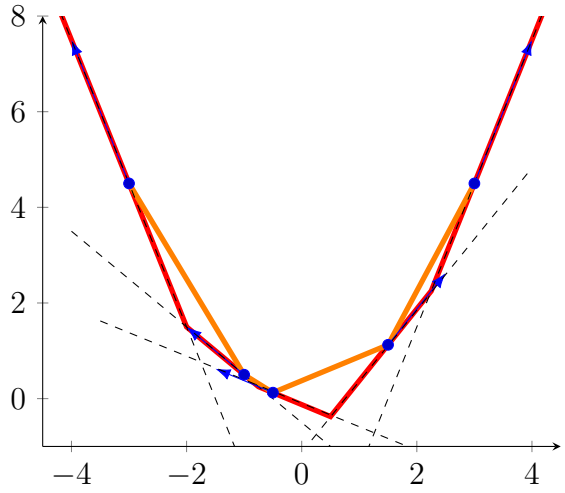
We consider the one-dimensional case, $n = 1$. Let us imagine that we run a first-order algorithm (that is, an algorithm which can access the value of f and ∇f at any desired point, and must return an approximate minimizer based on this information only). After some time, the algorithm has queried the values of f and ∇f at several points, for instance $\left\{-3, -1, -\frac{1}{2}, \frac{3}{2}, 3\right\}$. The gathered information is represented on the following picture.



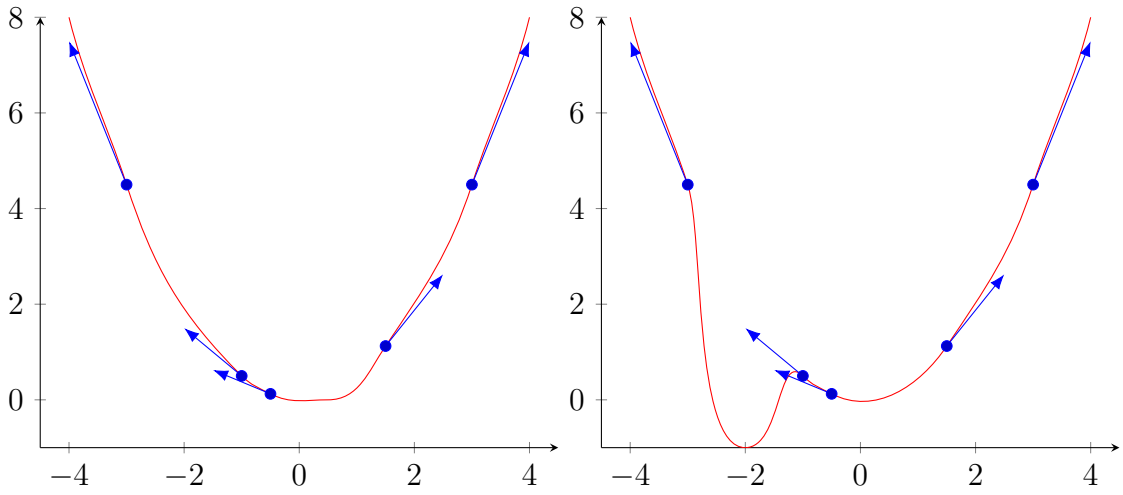
If f is convex, this already gives significant information on the minimum and minimizer of f. Indeed, the graph of f is above its tangents, and below its chords, which provides upper and lower bounds for f. In the specific example considered here, these upper and lower bounds are respectively represented in orange and red on the following picture. One can use them to deduce the following two properties :

1. The minimum of f is between $-3/8$ and $1/8$.
2. The minimizer(s) of f belong(s) to the interval $[-1/2; 5/6]$.

In particular, from this information, one knows the value of $\min f$ up to precision 0.5 and the minimizer up to precision 1.325.



But if f is not convex, this information does not allow to distinguish, for instance, the following two functions.



The function represented on the left reaches its minimum at $1/2$, and this minimum is 0. The function on the right reaches its minimum at -2, and this minimum is -1. The difference between the minimums of these two functions is 1, and the difference between the minimizers is 2.5 : One cannot produce estimations for the minimal value and minimizer of f comparable to the convex setting.

Intuitively, to compute a trustworthy approximation of $\min f$ or $\operatorname{argmin} f$ without the convexity assumption, one needs to sample f on a fine grid. As soon as there is a "hole" in the sampling set [1], one cannot know whether the function takes large or small values in this hole, hence one cannot compute a precise estimate of $\min f$ or $\operatorname{argmin} f$. In the one-dimensional case, it may be possible to sample f on a fine grid, but if n is large, this is out of question : The number of sampling points on a fine grid grows exponentially with the dimension.

As a consequence, if f is not convex, we must give up the idea of finding an approximate minimizer. In the rest of the lecture, we will see which kind of points we can hope to find, and how.

1. The *sampling set* is the set of points at which the algorithm queries the values of f and ∇f. In our example, it is $\{-3, -1, -1/2, 3/2, 3\}$.

2 Critical points

A first idea is to look for a *local minimizer* instead of a global one. It turns out that this is also out of reach, at least for pathological functions. Thus, we lower our expectations again : instead of looking for a local minimizer, we simply look for a point at which "the derivatives of f satisfy the same properties as at a local minimizer".

Proposition 2.1. *For any $x \in \mathbb{R}^n$, we denote* $\operatorname{Hess} f(x)$ *the Hessian of f at x, that is the $n \times n$ matrix whose (i,j)-th coefficient is $\frac{\partial^2 f}{\partial x_i \partial x_j}(x)$.*

If x is a local minimizer of f, then

$$\nabla f(x) = 0 \text{ and } \operatorname{Hess} f(x) \succeq 0.$$

Conversely, if $\nabla f(x) = 0$ and $\operatorname{Hess} f(x) \succ 0$*, then x is a local minimizer of f.*

Définition 2.2. *We say that an element x of $\mathbb{R}^n$ is*

- *a <u>first-order critical point</u> of f if $\nabla f(x) = 0$,*
- *a <u>second-order critical point</u> of f if $\nabla f(x) = 0$ and* $\operatorname{Hess} f(x) \succeq 0$*.*

Exemple 2.3. *We consider the map $f : (x_1, x_2) \in \mathbb{R}^2 \to x_1^2 - x_2^2 \in \mathbb{R}$.*

Its gradient and Hessian have the following formulas :

$$\forall x = (x_1, x_2) \in \mathbb{R}^2, \quad \nabla f(x) = (2x_1, -2x_2) \quad \text{and} \quad \operatorname{Hess} f(x) = \left(\begin{smallmatrix} 2 & 0 \\ 0 & -2 \end{smallmatrix}\right).$$

Therefore, f has a single first-order critical point, which is $(0,0)$. This point is not a second-order critical point, because $\left(\begin{smallmatrix} 2 & 0 \\ 0 & -2 \end{smallmatrix}\right)$ is not semidefinite positive.

Although second-order critical points are not always local minimizers [2], the two notions nevertheless coincide for many functions f. In addition, one numerically observes that, in various interesting situations (including the training of neural networks), f has no local minimizer other than its global one, or, at least, all its local minimizers are approximate global minimizers (meaning $f(x) \approx f(x_*)$). It is thus of practical importance to be able to find critical points.

2. The map $(x \to x^3)$ has a second-order critical point at 0, but no local minimizer.

3 Convergence of gradient descent

Let us first consider the simplest first-order algorithm, gradient descent. We have studied, in a previous lecture, its convergence properties in the convex setting. How does it perform in the non-convex one ?

We assume that f is L-smooth for some $L > 0$: For any $x, y \in \mathbb{R}^n$,

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||.$$

We consider gradient descent with constant stepsize, equal to $1/L$: starting from an arbitrary $x_0 \in \mathbb{R}^n$, we define a sequence $(x_t)_{t\in\mathbb{N}}$ by

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

3.1 Convergence to a first-order critical point

Théorème 3.1. *Let $T \in \mathbb{N}$ be fixed. We consider the following algorithm :*

1. *Run T steps of gradient descent, which defines a sequence $(x_0, x_1, \ldots, x_T)$.*
2. *Compute $T_{min} = \text{argmin}_{0\leq t\leq T} ||\nabla f(x_t)||$ and define $\tilde{x}_T = x_{T_{min}}$.*
3. *Return $\tilde{x}_T$.*

Then

$$||\nabla f(\tilde{x}_T)|| \leq \sqrt{\frac{2L(f(x_0) - f(x_*))}{T}}.$$

We say that $\tilde{x}_T$ is a $O(1/\sqrt{T})$-approximate first-order critical point.

Démonstration. Because f is L-smooth, it holds

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

Consequently,

$$\begin{aligned}\sum_{t=0}^{T-1} ||\nabla f(x_t)||^2 &\leq 2L \sum_{t=0}^{T-1} f(x_t) - f(x_{t+1}) \\ &= 2L(f(x_0) - f(x_T)) \\ &\leq 2L(f(x_0) - f(x_*)).\end{aligned}$$

Since $||\nabla f(\tilde{x}_T)|| \leq ||\nabla f(x_t)||$ for any $t \leq T$,

$$T||\nabla f(\tilde{x}_T)||^2 \leq 2L(f(x_0) - f(x_*)),$$

which implies

$$||\nabla f(\tilde{x}_T)|| \leq \sqrt{\frac{2L(f(x_0) - f(x_*))}{T}}.$$

□

3.2 Convergence to a second-order critical point

The previous theorem shows that gradient descent allows to find approximate first-order critical points, and even provides a convergence rate. For second-order critical points, the picture is more complicated.

For some choices of initial points x_0, it may happen that gradient descent does not get close to an approximate second-order critical point, even when run for an infinite number of steps. For instance, if x_0 is a first-order critical point of f, but not a second-order critical point, then

$$x_0 = x_1 = x_2 = \dots,$$

because $\nabla f(x_0) = 0$, hence gradient descent stays stuck at x_0 and never reaches a second-order critical point.

Théorème 3.2 (Lee, Simchowitz, Jordan, Recht 2016)**.** *We assume that*

- *f has only a finite number of first-order critical points ;*
- *for any $M \in \mathbb{R}$, $\{x \in \mathbb{R}^n, f(x) \leq M\}$ is bounded.*

We consider gradient descent with constant stepsize $\alpha \in]0; \frac{1}{L}[$.

For almost any x_0 (that is, for all x_0 outside a zero-Lebesgue measure set), $(x_t)_{t\in\mathbb{N}}$ converges to a second-order critical point.

Intuition of proof. The finiteness of the critical set and the boundedness of the level sets of f imply that $(x_t)_{t\in\mathbb{N}}$ converges to a first-order critical point whatever x_0. We admit this fact for simplicity.

We must show that, if x_{crit} is a first-order but not a second-order critical point of f, then $(x_t)_{t\in\mathbb{N}}$ does not converge to x_{crit}, for almost any x_0. We consider such a critical point ; up to translation, we can assume that it is 0.

We make the (very) simplifying hypothesis that f is quadratic in a ball centered at 0, whose radius we call r_0 :

$$\forall x \in B(0, r_0), \quad f(x) = \frac{1}{2} \langle x, Mx \rangle + \langle x, b \rangle ,$$

for some $n \times n$ symmetric matrix M.

For any $x \in B(0, r_0)$, $\nabla f(x) = Mx + b$. Since 0 is a first-order critical point, we necessarily have $b = 0$. In addition, $\operatorname{Hess} f(x) = M$ for any $x \in B(0, r_0)$: for all i, j,

$$\begin{aligned}
(\operatorname{Hess} f(x))_{i,j} &= \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \\
&= \frac{\partial^2}{\partial x_i \partial x_j} \left(x \to \frac{1}{2} \sum_{1 \leq k,l \leq n} M_{kl} x_k x_l \right)(x) \\
&= \frac{1}{2} \sum_{1 \leq k,l \leq n} M_{k,l} \frac{\partial^2}{\partial x_i \partial x_j} \left(x \to x_k x_l \right)(x) \\
&= \frac{1}{2} \sum_{1 \leq k,l \leq n} M_{k,l}. \begin{cases} 0 & \text{if } \{k,l\} \neq \{i,j\} \\ 1 & \text{if } i \neq j \text{ and } (k,l) = (i,j) \text{ or } (j,i) \\ 2 & \text{if } i = j = k = l \end{cases} \\
&= \frac{1}{2} \begin{cases} (M_{i,j} + M_{j,i}) & \text{if } i \neq j \\ 2M_{i,i} & \text{if } i = j \end{cases} \\
&= M_{i,j}.
\end{aligned}$$

The assumption that 0 is not a second-order critical point is then equivalent to the fact that $M \not\geq 0$.

As explained in the lecture of October 8, we can assume, up to an orthogonal change of coordinates, that M is diagonal :

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

with $\lambda_1 \geq \cdots \geq \lambda_n$ the eigenvalues of M. Since $M \not\geq 0$, at least the smallest eigenvalue of M is negative : $\lambda_n < 0$.

If the sequence $(x_t)_{t \in \mathbb{N}}$ of gradient descent iterates converges to 0, then

x_t belongs to $B(0, r_0)$ for any t large enough, in which case

$$\begin{aligned} x_{t+1} &= x_t - \alpha \nabla f(x_t) \\ &= x_t - \alpha M x_t \\ &= \begin{pmatrix} (1-\alpha\lambda_1)x_{t,1} \\ \vdots \\ (1-\alpha\lambda_n)x_{t,n} \end{pmatrix}. \end{aligned}$$

We fix t_0 such that this relation holds for any $t \geq t_0$. Then, for any $s \in \mathbb{N}$,

$$x_{t_0+s} = \begin{pmatrix} (1-\alpha\lambda_1)^s x_{t_0,1} \\ \vdots \\ (1-\alpha\lambda_n)^s x_{t_0,n} \end{pmatrix}.$$

If the sequence converges to 0, all the coordinates of x_{t_0+s} must go to 0 when s goes to $+\infty$ (for any fixed t), which means that

$$\forall k \in \{1, \dots, n\}, \quad (1 - \alpha\lambda_k)^s x_{t_0,k} \overset{s \to +\infty}{\to} 0. \tag{1}$$

We have said that $\lambda_n < 0$, hence $1 < 1 - \alpha\lambda_n$ and $(1 - \alpha\lambda_n)^s \not\to 0$ when $s \to +\infty$. In order for Property (1) to hold, we must therefore have

$$x_{t_0,n} = 0.$$

To summarize, we have shown that, if $(x_t)_{t\in\mathbb{N}}$ converges to 0, then, for some t_0,

$$x_{t_0} \in \mathcal{E} \overset{def}{=} \{z \in B(0, r_0) \text{ such that } z_n = 0\}.$$

As a consequence,

$$x_0 \in (\mathrm{Id} - \alpha\nabla f)^{-t_0}(\mathcal{E}).$$

(For any map $g : \mathbb{R}^n \to \mathbb{R}^n$, we define $g^{-t_0}(\mathcal{E})$ as the set of points x such that $g^{t_0}(x) = g \circ \cdots \circ g(x) \in \mathcal{E}$.) Therefore, the set of initial points x_0 for which the gradient descent iterates may converge to 0 is included in

$$\bigcup_{t\in\mathbb{N}} (\mathrm{Id} - \alpha\nabla f)^{-t}(\mathcal{E}).$$

The set $\mathcal{E}$ has zero Lebesgue measure and one can check that $\mathrm{Id} - \alpha\nabla f$ is a diffeomorphism, hence $(\mathrm{Id} - \alpha\nabla f)^{-t}(\mathcal{E})$ has zero Lebesgue measure for any $t \in \mathbb{N}$, and the set of "problematic" initial points also has zero Lebesgue measure.

□

4 A second-order method

The theorem stated in the previous paragraph only states that gradient descent reaches a second-order critical point "in the limit" (for almost any initial point x_0). It does not provide complexity estimates. This is unavoidable : in high dimension, gradient descent may converge extremely slowly in the worst case.

To overcome this possible slow convergence, several strategies are possible. One of them is to add "noise" to gradient iterates from time to time, to help them get away faster from first-order critical points. The interested reader will find a description in How to escape saddle points efficiently, by C. Jin, R. Ge, P. Netrapalli, S. Kakade and M. Jordan (ICML 2017)

Another one is to explicitely exploit the information provided by second-order derivatives. This yields the family of *second-order methods*. In this section, we briefly describe one member of this family : the trust-region method.

The starting point is that, in the same way that ∇f provides a linear approximation of f around any point, Hess f provides a (more precise) quadratic approximation.

Proposition 4.1. *For any $x \in \mathbb{R}^n$,*

$$f(x+h) = f(x) + \langle h, \nabla f(x)\rangle + \frac{1}{2}\langle h, \text{Hess}\, f(x)h\rangle + o(||h||^2).$$

To define x_{t+1} from x_t, it is therefore reasonable to set

$$h_t = \underset{||h||\leq R_t}{\text{argmin}} \left(f(x) + \langle h, \nabla f(x)\rangle + \frac{1}{2}\langle h, \text{Hess}\, f(x)h\rangle \right)$$

and $x_{t+1} = x_t + h_t$. (In the definition of h_t, R_t is a positive number, the *trust radius*, whose choice is important for the good behavior of the algorithm.)

We provide convergence guarantees for this algorithm under the assumption that Hess f is L_2-Lipschitz for some $L_2 > 0$:

$$\forall x, y, h \in \mathbb{R}^n, \quad ||(\text{Hess}\, f(x) - \text{Hess}\, f(y))h|| \leq L_2||x-y||\,||h||.$$

Théorème 4.2. *Let $\epsilon > 0$ be fixed.*

We run the trust-region algorithm as described above, with $R_t = \frac{\sqrt{\epsilon}}{L_2}$ for any t. We stop the algorithm if

$$\frac{||\nabla f(x_t) + \text{Hess}\, f(x_t)h_t||}{||h_t||} \leq \sqrt{\epsilon}$$

and return x_{t+1}.

For any $x_0 \in \mathbb{R}^n$*, the algorithm stops after at most* $O\left(\frac{L_2^2(f(x_0)-f(x_*))}{\epsilon^{3/2}}\right)$ *iterations and the output* x_{final} *is an approximate second-order critical point, in the sense that*

$$||\nabla f(x_{final})|| \lesssim \frac{\epsilon}{L_2} \quad \textit{and} \quad \lambda_{\min}\left(\operatorname{Hess} f(x_{final})\right) \gtrsim -\sqrt{\epsilon}.$$

(The notation "$\lesssim$*" means "smaller up to a moderate multiplicative constant" and* λ_{min} *is the smallest eigenvalue.)*

5 References

- Gradient descent only converges to minimizers, by J. D. Lee, M. Simchowitz, M. Jordan and B. Recht, in the Conference on Learning Theory (COLT), 2016.
- Second order optimization algorithms I, lecture notes by Y. Ye, available at http://web.stanford.edu/class/msande311/2017lecture13.pdf.
- Computing a trust region step, by J. J. Moré and D. C. Sorensen, in the SIAM journal on scientific and statistical computing, volume 4, number 3, 1983.