M2RI UT3
S10 - Stochastic Optimization algorithms

S. Gadat

Toulouse School of Economics

Université Toulouse I Capitole

1er janvier 2017

Table des matières

Chapitre 1

Convex optimisation

We briefly present in this chapter some motivations around optimisation and statistics and then describe some gentle remainders on convex analysis.

1.1 Motivations from statistics

Machine learning is an academic field (and also a research field) that looks for efficient algorithms for estimating an unknown relationship between X (observed variables) and Y (variable that should be predicted) from a set of data $(X_1, Y_1), \dots, (X_n, Y_n)$.

The starting point of any method is the development of a credible model that links Y to X. In many applications, such link by no means is deterministic and the baseline assumption is the existence of a set of statistical models $(\mathbb{P}_\theta)_{\theta\in\Theta}$ such that the variables (X, Y) are distributed according to $\mathbb{P}_{\theta_0}$ where θ_0 is an unknown parameter in Θ. Instead of estimating the joint law of (X, Y), we are rather interested in the conditional distribution of Y given X and with a slight abuse of notation, $\mathbb{P}_{\theta_0}(|X)$ will represent the distribution of what we want to predict (the variable Y) given the value of θ_0 and the value of the observation X.

From a statistical point of view, it is needed to estimate θ_0 from the set of observations $(X_1, Y_1), \dots, (X_n, Y_n)$ and the common efficient way to produce such estimation relies on the likelihood of the observations given the value of θ.

In its full generality, this problem is difficult (too difficult from a statistical point of view) and it is necessary to impose some restrictions on the model generality to obtain feasible and trustly resolutions. Below, we provide a brief non-exhaustive list of problems.

1.1.1 Linear model

This paragraph is motivated by the simplest link that may exist between two continuous random variables X and Y. We assume that (X, Y) are linked with a **Gaussian** linear model :

$$\mathcal{L}(Y|X) = \mathcal{N}\left(\langle\theta_0, X\rangle, \sigma^2\right),$$

where $X \in \mathbb{R}^p$, $\theta_0 \in \mathbb{R}^p$ is the unknown parameter and σ^2 is the known (or unknown) variance parameter. In that case, the likelihood of the observations may be written as

$$\forall\theta \in \mathbb{R}^p \qquad L(\theta) = \prod_{i=1}^{n} \frac{e^{-|Y_i - \langle\theta, X_i\rangle|^2/2\sigma^2}}{\sqrt{2\pi}\sigma}.$$

The statistical model being regular enough, the M.L.E. (maximum likelihood estimator) is an optimal statistical estimation procedure (asymptotically unbiased and with a minimal variance).

The M.L.E. $\hat{\theta}_n$ attains in particular the Cramer-Rao efficiency lower bound and is a maximum of $\ell : \theta \longmapsto \log L(\theta)$. Hence, finding $\hat{\theta}_n$ is equivalent to the minimization of $-\ell$ given by

$$-\ell(\theta) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} |Y_i - \langle \theta, X_i \rangle|^2 + n \log(\sqrt{2\pi}\sigma).$$

Assuming σ known, the minimization of $-\ell$ is then equivalent to the minimization of the classical sum of square criterion

$$\forall \theta \in \mathbb{R}^p \qquad U(\theta) := \sum_{i=1}^{n} |Y_i - \langle \theta, X_i \rangle|^2. \tag{1.1}$$

An explicit formula exists for the minimizer of U : if we denote $X = [X_1; \ldots; X_n]$ the design matrix of size $n \times p$ and Y the column vector of size $n \times 1$, then the M.L.E. is given by

$$\hat{\theta}_n := ({}^t XX)^{-1} \, {}^t XY. \tag{1.2}$$

Remark 1.1.1 *We should remark that Equation* (1.2) *is true as soon as the Fisher information matrix* $M = {}^t XX$ *is invertible. It corresponds to the situation where* U *given by Equation* (1.1) *is a* ***strongly convex function****.*

We will see in this chapter the exact meaning of this strong convexity, and some important consequences for the minimization of U.

1.1.2 Logistic regression

This paragraph is motivated by the simplest link that may exist between one continuous random variables X and a binary one Y. Hence, the problem belongs to the supervised classification framework : we observe X and want to predict the expected value of Y among $\{0, 1\}$. We assume that a hidden parameter $\theta_0 \in \Theta$ exists such that

$$\mathbb{P}(Y = 1 | X = x) = p(x, \theta_0).$$

If the observations are i.i.d., then the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \mathbb{P}(Y_i = 1 | X = X_i) = \prod_{i=1}^{n} p(X_i, \theta)^{Y_i} (1 - p(X_i, \theta))^{1 - Y_i} \tag{1.3}$$

If the probability of success of Y is independent of X, then $p(X, \theta_0) = p_0$ and the M.L.E. in the classification model is then estimated by :

$$\frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Now, if we assume an inhomogeneity in the relationship between Y and X, we then need to fix a set of constraints to produce an easy estimation of $\mathbb{P}(Y = 1|X)$. The natural statistical answer to this problem is to use an almost standard linear regression to fix some constraints on the function $(x, \theta) \longrightarrow p(x, \theta)$.

Unfortunately, the range of values of $\langle x, \theta \rangle$ is $[-\infty, \infty]$, which is not compatible with the range of values of p. Instead of describing p, we can imagine that this function describes $\log\left(\frac{p}{1-p}\right)$, for which the range of values is also $[-\infty, \infty]$. The logistic regression model is then defined by :

$$\log\left(\frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0|X = x)}\right) = \langle \theta_0, x \rangle. \tag{1.4}$$

An easy resolution leads to

$$\mathbb{P}(Y=1|X=x)=\frac{e^{\langle\theta_0,x\rangle}}{1+e^{\langle\theta_0,x\rangle}}.$$

We should again stress the fact that using logistic regression to predict class probabilities is a modeling choice , just like it is a modeling choice to predict quantitative variables with linear regression. By no means this model is the unique way to predict Y given X when Y is a binary variable.

According to (1.4) and the likelihood formultion (1.3), we then deduce that

$$\forall\theta\in\mathbb{R}^p \qquad \ell(\theta)=\sum_{i=1}^{n}-\log\left(1+e^{\langle\theta,X_i\rangle}\right)+\sum_{i=1}^{n}Y_i\langle\theta,X_i\rangle$$

It is easy to show that $\theta\longrightarrow\log(1+e^{\langle\theta,X_i\rangle})$ is a convex function, so that $\theta\longrightarrow-\log(1+e^{\langle\theta,X_i\rangle})$ is concave. Hence, maximizing ℓ is equivalent to minimize the ***convex function*** given by

$$U(\theta):=\sum_{i=1}^{n}\log\left(1+e^{\langle\theta,X_i\rangle}\right)-\sum_{i=1}^{n}Y_i\langle\theta,X_i\rangle. \tag{1.5}$$

In this case, we can immediately see that there is no explicit solution for the minimization of U, contrary to the explicit case of the linear regression. But we will see that it is still possible to exploit the convex property of U to obtain efficient algorithms for solving the logistic regression problem.

1.1.3 What goes wrong with Big Data ?

Main nowadays challenges in statistics and optimization are concerned by the large scale of the problems. In particular, there is a huge amount of informations for each observation (meaning that p is large, and maybe very large). Moreover, there is also a large number of observations : n is big too.

1.1.3.1 Unwell-posedness of some problems

Some consequences of this large number of variables and observations may be very annoying. The first one is concerned by a very important limitation when $p\geq n$. Concerning for example the case of the linear model defined in (1.1) and solved by (1.2), it is straightforward to check that the problem is not well posed : the squared matrix ${}^tXX$ is not invertible and all the nice theory of linear models has to be refreshed. Note that such a problem also occurs with logistic regression, and in many different area of statistics.

You will see how in another MSc course with Lasso estimator or Boosting algorithms...

1.1.3.2 Too much observations : on-line strategies

Another limitation is concerned by a too large number of observations available for a computer, that completely yields the saturation of its memory RAM. In such a case, we need to tackle the estimation problem by considering observations as sequential, and then only propose recursive strategies for estimation. We will then talk about **recursive methods**, in opposition with batch strategies. Many applications built from observations gathered by web browsers must be now sequential. We will see how we can handle these recursive arrivals to produce reliable statistical conclusions.

This will be the main topic of these Lecture Notes.

1.1.3.3 Interest of convex methods

The cornerstone of all these new challenges rely on the intensive use of convex analysis. Convex analysis makes it possible to produce easy numerical methods that are also efficient from a statistical point of view. Therefore, this chapter now starts the mathematical considerations with some remainders on convex analysis and convex algorithms.

1.2 General formulation of the optimization problem

1.2.1 Introduction

Minimization problem We consider $x \in \mathbb{R}^p$ (for statistical applications, it will be replaced by θ later on) and a real function $f : \mathbb{R}^p \longrightarrow \mathbb{R}$. We want to solve the problem :

$$(\mathcal{P}) \qquad \min_{x \in S} f(x),$$

where S is the set of constraints. We may have several type of minimization problems :

- Constrained problems : $S \subset \mathbb{R}^p$
- Unconstrained problems : $S = \mathbb{R}^p$
- Smooth problems : f is differentiable
- Nonsmooth problems : a part of the function f is not $\mathcal{C}^1$
- Linearly constrained problems : f is linear and $S \subset \mathbb{R}^p$
- Quadratic problem : f is quadratic w.r.t. x

Approximate solution - Cost efficiency In its full generality, a minimization problem is unsolvable, meaning that we do not have access to an explicit exact solution for the minimizer. Therefore, we will develop a theory that makes it possible to *approximate* solutions of $(\mathcal{P})$ with an accuracy $\epsilon > 0$. In particular, if $x^\star$ is the exact solution of $(\mathcal{P})$, then an ϵ solution x_ϵ satisfies :

$$\|f(x^\star) - f(x_\epsilon)\| \leq \epsilon.$$

Then, the efficiency of the method involves the numerical cost (that will depend on ϵ) needed to obtain an ϵ solution x_ϵ. Generally, the algorithms we will use are iterative, meaning that they compute an estimation of the ϵ solution recursively, and the numerical cost of the method is tightly linked to the number of iterations needed for a good approximation.

X'th order method To conclude the paragraph, let us mention that the standard formulation $(\mathcal{P})$ is called a functional model of optimization problems. Usually, for such models the standard assumptions are related to the smoothness of functional components. In accordance to degree of smoothness we can apply different types of oracle to obtain an optimization algorithm :

- Zero-order oracle : returns the value $f(x)$.
- First-order oracle : returns $f(x)$ and the gradient $\nabla f(x)$.
- Second-order oracle : returns $f(x), \nabla f(x)$ and the Hessian $D^2 f(x)$.

1.2.2 General bound for global optimization on Lipschitz classes

We consider a very poor setting almost "assumption free" on the minimization problem $(\mathcal{P})$. We assume that f is L-Lipschitz, defined by

Definition 1.2.1 (L-Lipschitz) *The objective function f is L-Lipschitz if and only if*

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

Such a function f is of course continuous, but not necessarily differentiable. Implicitely, we assumed that f is defined on $\mathbb{R}^n$ equipped with the sup norm $\|.\|_\infty$. This notation is simplified by $|.|$ below for the sake of convenience.

We now provide a very pessimistic bound on the numerical cost for solving an ϵ solution of $(\mathcal{P})$ when f has to be minimized on $\mathcal{B}_n$:

$$\mathcal{B}_n := \{x \in \mathbb{R}^n \,|\, \forall i \in \{1 \dots n\} : \quad 0 \leq x_i \leq 1\}.$$

Theorem 1.2.1 *An ϵ solution of $(\mathcal{P})$ with L Lipschitz function can be solved in*

$$\left\{\frac{L}{2\epsilon} + 2\right\}^n$$

operations.

Proof : The method then simply consists in defining an ϵ grid of $\mathcal{B}_n$, denoted by $\mathcal{G}_n$, and defined by

$$\mathcal{G}_n := \left\{x_{k_1,\dots,k_n} = (k_1\epsilon, \dots k_n\delta) \,|\, \forall i \in \{1, \dots, n\} : \quad k_i \in [\![0, \delta^{-1} \wedge 1]\!]\right\}.$$

Therefore, $\mathcal{G}_n$ is a regularly spaced grid on $\mathcal{B}_n$ with a window size equals to δ. It is straightforward to check that

$$|\mathcal{G}_n| \simeq \{\delta^{-1}\}^n$$

If we now want to find an ϵ solution of $(\mathcal{P})$, it is enough to compute f on $\mathcal{B}_n$ and locate its minimal value on the grid attained at x_δ. We then have

$$0 \leq f(x_\delta) - f(x^\star) \leq f(\tilde{x}) - f(x^\star),$$

where $x^\star$ is the closest point to $x^\star$ on the grid $\mathcal{G}_n$. Then, the Lipschitz assumption on L yields :

$$f(\tilde{x}) - f(x^\star) \leq L \times |\tilde{x} - x^\star| \leq L\frac{\delta}{2}.$$

With our goal to obtain an ϵ solution, we then need to choose

$$\delta \leq \frac{2\epsilon}{L}.$$

Then, the number of points in $\mathcal{G}_n$ is $(\delta + 2)^{-n}$, which concludes the proof. □

The result above justifies an upper complexity bound for our problem class. This result is quite informative, and is certainly based on a very poor method (an exhaustive search on a uniform grid). We still have some questions !

— Firstly, it may happen that our proof is too rough and the real performance of this algorithm is much better.

— Secondly, we still cannot be sure that the algorithm itself is a reasonable method for solving $(\mathcal{P})$. There may exist other schemes with much higher performance.

In order to answer these questions, we need to derive lower complexity bounds for the problem class. This is beyond the scope of these Lecture Notes. We will sometimes provide some well-known results in operation research that precise the complexity of a problem and the performance of the described numerical method related to the complexity lower bound.

For example, it can be shown the following "lower bound".

Theorem 1.2.2 *Assume that $\epsilon \leq \frac{L}{2}$, then solving an ϵ solution of $(\mathcal{P})$ with a zero-th order method is* ***at least***

$$\left\{\frac{L}{2\epsilon}\right\}^n.$$

1.2.3 Comments

Optimality Taken together, the complexity lower bound given by Theorem 1.2.2 and the "algorithm" described in Theorem 1.2.1 show that the real complexity of solving an ϵ solution of $(\mathcal{P})$ with zero-th order method on Lipschitz classes is exactly of the order $\left\{L\epsilon^{-1}\right\}^n$. Hence, the method described by the exhaustive search on the grid $\mathcal{G}_n$ is *optimal*.

Reasonnable computation Nevertheless, it can be rapidly seen that the above problem cannot be solved in a reasonnable time with suplementary assumptions. For example, take $L = 2$, $n = 10$ and $\epsilon = 10^{-2}$, the numerical cost requires 10^{20} operations ...

We should note, that the lower complexity bounds for problems with smooth functions, or for high-order methods are not much better than those of Theorem 1.2.2. This can be proved using an argument close to the original proof of Theorem 1.2.2.

Comparison of the above results with the upper bounds for NP-hard problems, which are considered as a classical example of very difficult problems in combinatorial optimization, is also quite disappointing. Hard combinatorial problems need 2^n operations only !

Beyond pessimistic bounds for big data problems We will introduce below a very desirable property for functions involved in $(\mathcal{P})$ that makes it possible to find an ϵ solution of $(\mathcal{P})$ much more efficiently ! In particular, we will pay a very specific attention on the effect of the **dimension** of the problem on the complexity of the proposed algorithm. Indeed, we plan to apply our optimization procedure to high dimensional problems involved in Big Data. Therefore, this preocupation is very legitimage !

1.3 Gradient descent

1.3.1 Differentiable functions

A first natural additional assumption on f concerns a smoothness property. We recall some elementary definitions on differential functions below.

Definition 1.3.1 (Differentiable function) *$f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is differentiable iff for any x in $\mathbb{R}^n$, a linear map ℓ_x exists such that*

$$\|f(x) - f(y) - \ell_x(y - x)\| = o(|x - y|).$$

Since a linear map can always be associated to a vector of $\mathbb{R}^n$ by duality, we can also define the *gradient* of f at point x as the unique vector of $\mathbb{R}^n$ such that

$$\|f(x) - f(y) - \langle \nabla f(x), y - x\rangle\| = o(|x - y|).$$

In standard situations, the gradient of f is the vector of partial derivatives of f with respect to each coordinates :

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x)\right).$$

As an example, we can compute the gradient vector of the function $f(x) = \frac{a_1}{2}x_1^2 + \ldots + \frac{a_n}{2}x_n^2$:

$$\nabla f(x) = (a_1 x_1, \ldots, a_n x_n).$$

If the function f is more complex, for example with interactions between variables, the expression of ∇f may be more complex. If $f(x) = x_1^2 + ax_1x_2 + bx_1 + x_2^2$, then

$$\nabla f(x) = (2x_1 + ax_2 + b, ax_1 + 2x_2).$$

1.3.2 Smoothness class and consequences

We define below the set of functions $\mathcal{C}^1_L$ that are differentiables with a L-Lipschitz gradient function :

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

This function class satisfies the following lemma.

Lemma 1.3.1 *For any $f \in \mathcal{C}^1_L$, we have*

$$\forall (x,y) \in \mathbb{R}^n \qquad |f(y) - f(x) - \langle \nabla f(x), y-x\rangle| \leq \frac{L}{2}\|y-x\|^2 \tag{1.6}$$

Proof : We use a first order Taylor expansion :

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \langle \nabla f(x+s(y-x)), y-x\rangle ds \\ &= f(x) + \langle \nabla f(x), y-x\rangle + \int_0^1 \langle \nabla f(x+s(y-x)) - \nabla f(x), y-x\rangle ds. \end{aligned}$$

The last term may be upper bounded and we get :

$$|f(y) - f(x) - \langle \nabla f(x), y-x\rangle| \leq \int_0^1 Ls\|y-x\|^2 ds = \frac{L}{2}\|y-x\|^2.$$

We then obtain the conclusion. □

The geometrical interpretation of the last inequality is very important, and is the baseline fact of many optimization methods. We define $x_0 \in \mathbb{R}^n$ and two quadratic functions given by :

$$\phi_\pm(x) = f(x_0) + \langle \nabla f(x_0), x - x_0\rangle \pm \frac{L}{2}\|x-x_0\|^2.$$

It is immediate to check that

$$\forall x \in \mathbb{R}^n \qquad \phi_-(x) \leq f(x) \leq \phi_+(x)$$

Therefore, if we want to minimize f, a reasonnable way to obtain an efficient algorithm is to approximate locally f by ϕ_+ and then use an explicit formula that minimizes ϕ_+.

We can even improve the approximation of f with a third order Taylor expansion (the proof is omitted).

Lemma 1.3.2 *If f is $\mathcal{C}^2$ with M Lipschitz Hessian, then*

$$\|\nabla f(y) - \nabla f(x) - D^2 f(x)(y-x)\| \leq \frac{M}{2}\|y-x\|^2,$$

and

$$\left| f(y) - f(x) - \langle \nabla f(x), y-x\rangle - \frac{1}{2}\langle D^2 f(x)(y-x), y-x\rangle \right| \leq \frac{M}{6}\|y-x\|^3$$

1.3.3 Gradient method

Now we are completely ready for studying the convergence rate of unconstrained minimization methods and in particular the simplest first order method, which is the gradient descent method. We provide below two reasons why such a method may be efficient.

1.3.3.1 Antigradient as the steepest descent

The baseline ingredient is represented by the fact that the antigradient at point x, given by $-\nabla f(x)$, is a direction of locally steepest descent of differentiable function. Since we are going to find its local minimum, the following scheme is the first to be tried :

Algorithm 1 Gradient descent scheme

InputFunction f. Stepsize sequences $(\gamma_k)_{k\in\mathbb{N}}$
Initialization : Pick $x_0 \in \mathbb{R}^n$.
Iterate

$$\forall k \in \mathbb{N} \qquad x_{k+1} = x_k - \gamma_k \nabla f(x_k). \tag{1.7}$$

Output : $\lim_{k\longrightarrow+\infty} x_k$

We will refer to this scheme as a **gradient method**. The scalar factor of the gradient, γ_k, is called the step size. Of course, it must be positive ! There are many variants of this method, which differ one from another by the step-size strategy. Let us consider the most important examples.

Definition 1.3.2 (Gradient descent - adaptive step-size) *The sequence $(\gamma_k)_{k\geq 0}$ is chosen in advance independently of f by :*

$$\gamma_k = \gamma > 0 \qquad \textit{constant step size},$$

or $(\gamma_k)_{k\geq 0}$ is decreasing :

$$\gamma_k = \frac{\gamma}{\sqrt{k+1}}.$$

Definition 1.3.3 (Gradient descent - line search backtracking) *The full relaxation step-size of the gradient descent is defined by :*

$$\gamma_k = \arg\min_{\gamma>0} f(x_k - \gamma \nabla f(x_k))$$

Among these strategies, we see that the first strategy is the simplest one. Indeed, it is often used, but mainly in the context of convex optimization. In that framework the behavior of functions is much more predictable than in the general nonlinear case (see below) and the gain brought by the gradient descent may be quantified easily.

The second strategy is completely theoretical, it is an abstract tool but it is never used in practice since even in one-dimensional cases we cannot find an exact minimum in finite time.

1.3.3.2 Gradient descent as a Maximization-Minimization method

The second way to understand the gradient descent method is more geometrical and relies on the understanding of « Maximization-Minimization »algorithm. The geometrical idea is illustrated in Figure 1.1.

Imagine that :

— we are able to produce for each point $y \in \mathbb{R}^n$ an auxiliary function $x \longrightarrow G(x, y)$ such that

$$\forall x \in \mathbb{R}^n \qquad f(x) \leq G(x, y) \qquad \text{and} \qquad f(y) = G(y, y).$$

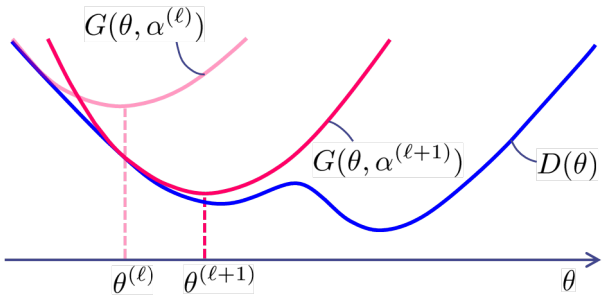

FIGURE 1.1: Geometrical illustration of the MM algorithm : we minimize $\theta \longrightarrow D(\theta)$ with the help of some auxiliary functions $\theta \longrightarrow G(\theta, \alpha)$.

— we have an **explicit exact formula** that makes it possible to **minimize** the auxiliary function $x \longrightarrow G(x, y)$:

$$\arg\min_{x \in \mathbb{R}^n} G(x, y)$$

Then, a possible method to minimize f seems to produce a sequence $(x_k)_{k \geq 0}$ as follows :

Algorithm 2 Maximization-Minimization algorithm

InputFunction f. Family of auxiliary functions G
Initialization : Pick $x_0 \in \mathbb{R}^n$.
Iterate Compute $G_k(x) := G(x, x_k)$ and solve

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} G_k(x). \tag{1.8}$$

Output : $\lim_{k \longrightarrow +\infty} x_k$

The keypoint is that we have with Lemma 1.3.1 a function ϕ_+ that is an upper bound of f :

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

and ϕ_+ is easy to minimize : it is a second degree polynomial in x ! In particular, we can check that ϕ_+ is convex and coercive :

$$\lim_{\|x\| \longrightarrow +\infty} \phi_+(x) = +\infty.$$

Moreover, the gradient of ϕ_+ can be computed :

$$\nabla \phi_+(x) = \nabla f(y) + L(x - y).$$

Therefore, the minimizer of ϕ_+ is

$$\arg\min_{x \in \mathbb{R}^n} \phi_+(x) = y - \frac{1}{L} \nabla f(y).$$

We then conclude that the MM algorithm associated to the **surrogate auxiliary function** ϕ_+ is nothing more than the standard gradient descent with a step-size L^{-1}.

1.3.3.3 Theoretic guarantee

We consider the first gradient descent with constant step size γ.

Theorem 1.3.1 *Let f be a positive $\mathcal{C}^1_L$ function, then the gradient descent method applied with $\gamma^\star = L^{-1}$ satisfies*

$$\lim_{k \longrightarrow +\infty} \nabla f(x_k) = 0.$$

Proof : First step : Optimization of the gradient descent We define $x = x_k$ and $y = x_{k+1}$ and aim to make $f(y)$ as small as possible starting from x in one step. In this view, we use the bound given by Inequality (1.6). We know that

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

Considering $y = x - \gamma \nabla f(x)$, we have

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

where we have used the equality $\|y - x\| = \gamma \|\nabla f(x)\|$. We therefore obtained :

$$f(y) \leq f(x) - \gamma \left(1 - \frac{\gamma L}{2}\right) \|\nabla f(x)\|^2.$$

Our strategy is then to minimize $f(y)$ in one step, meaning that we are looking for γ such that $\gamma \left(1 - \frac{\gamma L}{2}\right)$ is as large as possible. The function of γ above attains its maximal value at $\gamma^\star = L^{-1}$. With this simple gradient descent scheme, we then obtain the inequality (known as a **descent** inequality) :

$$f(x - L^{-1} \nabla f(x) \leq f(x) - \frac{1}{2L} \|\nabla f(x)\|^2. \tag{1.9}$$

Second step : Convergence of the gradient descent We consider the sequence $(x_k)_{k \geq 1}$ defined by (1.7) with $\gamma_k = \gamma^\star = L^{-1}$. Inequality (1.14) shows that

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

We can sum all these inequalities between 0 and n and obtain that

$$f(x_{n+1}) - f(x_0) \leq -\frac{1}{2L} \sum_{k=0}^{n} \|\nabla f(x_k)\|^2.$$

This may be rewritten as

$$\sum_{k=0}^{n} \|\nabla f(x_k)\|^2 \leq 2L \left[f(x_0) - f(x_{n+1})\right] \leq 2L[f(x_0) - f(x^\star)].$$

We then conclude that

$$\lim_{k \longrightarrow +\infty} \nabla f(x_k) = 0,$$

meaning that the sequence $(x_k)_{k \geq 0}$ converges owards the set of critical points of f. We should note that at the moment, we do not know anything on the nature of this critical point, nor on the pointwise convergence of the sequence $(x_k)_{k \geq 0}$.

However, if we define

$$g_n^\star = \min_{0\leq k\leq n} \|\nabla f(x_k)\|,$$

we then obtain

$$g_n^\star \leq \frac{\sqrt{2L[f(x_0) - f(x^\star)]}}{\sqrt{n+1}}. \tag{1.10}$$

□

1.3.4 Rate of convergence of Gradient Descent

We can now make more precise the ability of Gradient descent to minimize a smooth function f and then solve $(\mathcal{P})$. Inequality (1.10) quantifies a **rate of convergence** of the method. Indeed, if the Hessian of f is not degenerated around $x^\star$, then f may be approximated by

$$f(x) = f(x^\star) + \langle D^2 f(x^\star)(x - x^\star), (x - x^\star)\rangle + o(\|x - x^\star\|^2),$$

and

$$\|\nabla f(x)\| = \|D^2 f(x^\star)(x - x^\star)\|.$$

It means that $\|\nabla f(x)\|$ quantifies the distance between x and $x^\star$. Consequently, if we are looking for an approximate ϵ solution of $(\mathcal{P})$, then it is enough to compute a solution x_n such that

$$\|\nabla f(x_n)\| \lesssim \epsilon.$$

With a gradient descent scheme applied with $\gamma = \gamma^\star$, if we want to find an ϵ approximate solution of $(\mathcal{P})$, we need to run n step such that

$$g_n^\star \leq \epsilon.$$

Applying Inequality (1.10), we deduce that the integer n should be chosen such that

$$n_\epsilon \geq 2L\epsilon^{-2}[f(x_0) - f(x^\star)].$$

We immediately remark that the rate is greatly improved (comparing to the ϵ^{-d} obtained in the general L-Lipschitz minimization problem) : the dimension of the problem does not seem to appear in the complexity of the problem since whatever the dimension d is, the amount of iteration is proportionnal to $\epsilon^{-\mathbf{2}}$.

Moreover, the larger the difference between $f(x_0)$ and $f(x^\star)$, the longer the computation needed to find an ϵ approximate solution of $(\mathcal{P})$.

However, in its full generality, the gradient descent scheme suffers from two important drawbacks. First, we only know that $\nabla f(x_n) \longrightarrow 0$ as $n \longrightarrow +\infty$. Indeed, nothing is known about the nature of the limit. In particular, the limit can also be a local maxima of f. Second, even though the limit of the sequence is a minima of f, we do not know if the limit is a global minima or a local trap. The goal of the next paragraph is to enrich the functional space with a very desirable property : convexity.

1.4 Convexity

In this section we deal with the unconstrained minimization problem $(\mathcal{P})$ where the function f is smooth enough. In the previous paragraph, we were trying to solve this problem under very weak assumptions on function f. And we have seen that in this general situation we cannot

do too much : impossible to guarantee convergence even to a local minimum, impossible to get acceptable bounds on the global performance of minimization schemes. Below, we introduce some reasonable assumptions on function / to make our problem more tractable.

1.4.1 Definition of convex functions

For that, let us try to determine the desired properties of a class of differentiate functions we want to work with : the set of convex functions.

Definition 1.4.1 *A function $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ is convex iff*

$$\forall (x,y) \in \mathbb{R}^p \quad \forall \lambda \in [0,1] \qquad f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y). \tag{1.11}$$

We left to the reader the proof of the following obvious facts :

Proposition 1.4.1 *The next three points hold.*
i) The sum of two convex functions is convex.
ii) Affine functions are convex.
iii) If f is convex and ϕ is an affine function, then $f \circ \phi$ is convex.

We will oftenly use the particular case of smooth differentiable convex functions. In that case, the definition above may be translated into a more tractable characterisation.

Proposition 1.4.2 *If f is $\mathcal{C}^1$ differentiable from $\mathbb{R}^n$ to $\mathbb{R}$, then f is convex if and only if*

$$\forall (x,y) \in \mathbb{R}^n \qquad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle. \tag{1.12}$$

Proof : We consider $(x,y) \in \mathbb{R}^n$ and $\lambda \in [0,1]$ and define

$$h(\lambda) = f(\lambda y + (1-\lambda)x) - \lambda f(y) - (1-\lambda) f(x).$$

We have

$$h'(\lambda) = \frac{d}{d\lambda} f(x + \lambda(y-x)) - (f(y) - f(x)) = \langle \nabla f(x), (y-x) \rangle + f(x) - f(y).$$

Then, we can check that if (1.12) holds, then h is a decreasing function. Since $h(0) = 0$, we can deduce that $h(\lambda) \leq 0$ for any value of $\lambda \in [0,1]$. Hence, f is convex and satisfies (1.11).

Conversely, we assume that f is convex so that $h(\lambda) \leq 0$ for any $\lambda \in [0,1]$. Since $h(0) = 0$, we deduce that $h'(0) \leq 0$, which is exactly Equation (1.12). □

1.4.2 Local minima of convex functions

A first important fact is stated below when we handle convex functions :

Theorem 1.4.1 *If f is convex and $\nabla f(x^\star) = 0$, then $x^\star$ is the **global** minimum of f.*

Proof : The proof is obvious when using Equation (1.12). We consider $y \in \mathbb{R}^n$ and see that

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

Hence, $f(x^\star)$ is the minimal value of f over $\mathbb{R}^n$.

Roughly speaking, for a convex function, being a critical point, or a local minima is equivalent to being a global minima.

1.4.3 Twice differentiable convex functions

We can obtain an even more tractable characterisation of a convex function f when f is twice differentiable. First, we prove the preliminary proposition.

Proposition 1.4.3 *A continuously differentiable function f is convex if and only if for any $(x, y) \in \mathbb{R}^n$, we have*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle > 0. \tag{1.13}$$

Proof : We assume f is convex. Then, we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \qquad \text{and} \qquad f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle.$$

Adding these two inequalities yields the conclusion of the first implication.

Conversely, we assume that (1.13) and take $(x, y) \in \mathbb{R}^n$. Then, we denote $x_s = x + s(y - x)$ and

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \langle \nabla f(x + s(y - x)), y - x \rangle ds \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x_s) - \nabla f(x), y - x \rangle ds \\ &= f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \frac{1}{s} \langle \nabla f(x_s) - \nabla f(x), x_s \rangle ds \\ &\geq f(x) + \langle \nabla f(x), y - x \rangle. \end{aligned}$$

□

We then obtain the main result of the paragraph.

Theorem 1.4.2 *A twice differentiable function f is convex if and only if $D^2 f \geq 0$.*

Proof : Let f a convex function and denote $x_s = x + sy$. Then, in view of (1.13), we have

$$0 \leq \frac{1}{s} \langle \nabla f(x_s) - \nabla f(x), x_s - x \rangle = \frac{1}{s} \langle \nabla f(x_s) - \nabla f(x), y \rangle = \frac{1}{s} \int_0^s \langle D^2 f(x + \lambda y)(y), y \rangle d\lambda$$

Now, taking the limit $s \longrightarrow 0$, we deduce that

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n \qquad \langle D^2 f(x)(y), y \rangle \geq 0,$$

meaning that $D^2 f(x)$ is a positive quadratic form. To obtain the reverse implication, we use a second order Taylor expansion :

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \int_0^1 \int_0^s \langle D^2 f(x + \lambda(y - x))(y - x), y - x \rangle d\lambda ds \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

This ends the proof. □

1.4.4 Examples

We describe below several examples of convex functions that are commonly encountered in statistics.

1. An affine function is convex :

$$\forall x \in \mathbb{R}^n \qquad f(x) = a + \langle b, x \rangle$$

2. A quadratic function is convex when the Hessian is symmetric and positive :

$$f(x) = a + \langle b, x\rangle + \frac{1}{2}^t xAx.$$

The Hessian of f is A, meaning that f is convex if and only if $A \geq 0$ and symmetric.

3. The exponential map is convex

$$\forall x \in \mathbb{R} \qquad f(x) = e^x \qquad f"(x) = e^x.$$

This is also true for the ℓ^p norm in $\mathbb{R}^n$ when $p \geq 1$:

$$\forall x \in \mathbb{R}^n \qquad f(x) = \|x\|_p = \left(\sum_{i=1}^n |x|^p\right)^{1/p}.$$

We should keep in mind this typical example, which will be the most important for us in this course.

4. The entropy function h is convex :

$$\forall x \in \mathbb{R}^n \qquad h(x) = x \log x.$$

This convexity property may be extended to the simplex of probability measures :

$$\forall p \in \mathcal{S}_n \qquad h(p) = \sum_{i=1}^n p_i \log p_i$$

The proofs of the later convexity properties are almost all immediate with the help of the Hessian criterion. The only non-trivial point concerns the convexity of the ℓ^p norm. Indeed, it is known as the Minkowski inequality (triangle inequality for ℓ^p norms) :

$$\begin{array}{lrcl} \forall (x,y) \in \mathbb{R}^n \quad \forall \lambda \in [0,1] & f(\lambda x + (1-\lambda)y) & = & \|\lambda x + (1-\lambda)y\|_p \\ & & \leq & \|\lambda x\|_p + \|(1-\lambda)y\|_p \\ & & = & \lambda\|x\|_p + (1-\lambda)\|y\|_p. \end{array}$$

1.4.5 Minimization lower bound

Below, we provide a theoretical result that will describe a *lower bound* of the complexity for solving $(\mathcal{P})$ when f is convex. We aim to solve the following problem :

$$\text{Find } x \text{ s.t.} f(x) - f(x^\star) \leq \epsilon \qquad \text{where} \qquad f \text{ is convex with } L \text{ Lipschitz gradient.}$$

Implicitely, providing a lower bound of the complexity of the above problem requires to build the *worst* function. Hence, the underlying construction is more or less artificial and only interesting for the upper bound we will obtain below.

Theorem 1.4.3 *For any $k \in \mathbb{N}$ and any $x_0 \in \mathbb{R}^n$, a function f exists with L Lipschitz gradient such that for any first order method :*

$$f(x_k) - f(x^\star) \geq \frac{3L\|x_0 - x^\star\|^2}{32(k+1)^2}.$$

Theorem 1.4.3 provides a lower bound for the complexity of any first order method that permit to obtain an ϵ approximation of the minimum of f. In particular, we can check that we need at least $k \geq \epsilon^{-1/2}$ iteration of a first order method to obtain an ϵ approximation. A priori, this is much more smaller than the gradient descent complexity obtained before with ϵ^{-2} iterations needed (without any convexity assumption). The proof of such a result may be found in the Lectures Notes of Nesterov.

In the next Section, we will provide an optimal method for minimizing in such convex classes. However, we will also introduce another class of function where the minimization is much more easier (from a numerical point of view).

1.5 Minimization of convex functions

We recall that a continuously differentiable function f is L-smooth if the gradient ∇f is L-Lipschitz, that is

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

Note that if f is twice differentiable then this is equivalent to the eigen-values of the Hessians being smaller than L. In this section we explore potential improvements in the rate of convergence under such a smoothness assumption. In order to avoid technicalities we consider first the unconstrained situation, where f is a convex and L-smooth function on $\mathbb{R}^n$. The next theorem shows that gradient descent, which iterates

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

attains a much faster rate in this convex situation than in the non-convex case given in Theorem 1.3.1 (the rate we obtained was $t^{-1/2}$).

Theorem 1.5.1 *Let f be convex and L-smooth on $\mathbb{R}^n$. Then gradient descent with $\eta = L^{-1}$ satisfies*

$$f(x_t) - f(x^\star) \leq \frac{2L\|x_0 - x^\star\|^2}{t - 1}$$

Before embarking on the proof we state a few properties of smooth convex functions. First, we recall the basic descent inequality :

$$0 \leq f(x) - f(y) - \langle \nabla f(y), x - y \rangle \leq \frac{L}{2}\|x - y\|^2, \tag{1.14}$$

which implies that :

$$f\left(x - \frac{1}{L}\nabla f(x)\right) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|^2.$$

Lemma 1.5.1 *Let f be such that Equation (1.14) holds true. Then for any $x, y \in \mathbb{R}^n$, one has :*

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

Proof : We define $z = y - \frac{1}{L}\left[\nabla f(y) - \nabla f(x)\right]$. We apply Inequality (1.14) and get

$$0 \leq f(z) - f(x) - \langle \nabla f(x), z - x \rangle,$$

leading to

$$f(x) - f(z) \leq \langle \nabla f(x), x - z \rangle.$$

Moreover, Inequality (1.14) also yields

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

Adding the two terms leads to

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

□

We can now prove Theorem 1.5.1.
Proof :
First step : convex inequalities. First, we apply the right hand side of Inequality (1.14) and obtain

$$f(x_{s+1}) - f(x_s) \leq -\frac{1}{2L}\|x_{s+1} - x_s\|^2.$$

In particular, if we denote by δ_s :

$$\delta_s := f(x_s) - f(x^\star),$$

we obtain

$$\delta_{s+1} \leq \delta_s - \frac{1}{2L}\|\nabla f(x_s)\|^2. \tag{1.15}$$

We can also apply the left hand side of Inequality (1.14) to obtain

$$\delta_s = f(x_s) - f(x^\star) \leq \langle \nabla f(x_s), x_s - x^\star \rangle \leq \|x_s - x^\star\|.\|\nabla f(x_s)\|, \tag{1.16}$$

where the last inequality is obtained with the Cauchy-Schwarz inequality.

Second step : $\|x_s - x^\star\|$ is a decreasing sequence. To show this result, we will use Lemma 1.5.1, which implies

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

We use this as follows :

$$\begin{aligned} \|x_{s+1} - x^\star\|^2 &= \left\| x_s - \frac{1}{L} - x^\star \right\|^2 \\ &= \|x_s - x^\star\|^2 - \frac{2}{L}\langle \nabla f(x_s), x_s - x^\star \rangle + \frac{1}{L^2}\|\nabla f(x_s)\|^2 \\ &\leq \|x_s - x^\star\|^2 - \frac{1}{L^2}\|\nabla f(x_s)\|^2 \end{aligned}$$

Therefore, $(\|x_s - x^\star\|)_{s \geq 1}$ is a decreasing sequence.
Third step : Convergence rate of $(\delta_s)_{s \geq 1}$. We know from (1.16) and the fact that $(\|x_s - x^\star\|)_{s \geq 1}$ is a decreasing sequence that

$$\frac{\delta_s}{\|x_1 - x^\star\|} \leq \frac{\delta_s}{\|x_s - x^\star\|} \leq \|\nabla f(x_s)\|.$$

Therefore, Equation (1.16) yields

$$\delta_{s+1} \leq \delta_s - \frac{1}{2L\|x_1 - x^\star\|^2}\delta_s^2.$$

If we define $\omega = 2L\|x_1 - x^\star\|^2$, we then obtain

$$\omega\delta_s^2 + \delta_{s+1} \leq \delta_s.$$

Dividing the above inequality by $\delta_s\delta_{s+1}$, we obtain

$$\omega\frac{\delta_s}{\delta_{s+1}} + \frac{1}{\delta_s} \leq \frac{1}{\delta_{s+1}}.$$

It implies that

$$\frac{1}{\delta_{s+1}} - \frac{1}{\delta_s} \geq \omega\frac{\delta_s}{\delta_{s+1}} \geq \omega.$$

Summing these inequalities from 1 to $n-1$, we obtain

$$\frac{1}{\delta_n} \geq \omega(n-1).$$

We then obtain

$$f(x_n) - f(x^\star) \leq \frac{2L}{n-1}\|x_0 - x^\star\|^2$$

□

Let us briefly discuss on the way the proof works. The main ingredient is a Lyapunov function that traduces a reverting effect from x_s to x_{s+1}. This function, in this case, is simply f itself since we then obtain :

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

Note that the second step of the proof also shows that $\|x - x^\star\|^2$ is also a second informative Lyapunov function. Then, some algebraic tricky relationships permit to conclude a rate of convergence for $(x_s)_{s\geq 1}$. This rate is still polynomial in our convex case, but we see below that this rate is greatly improve as soon as strong convex properties hold.

1.6 Strong convexity

1.6.1 Definition

Definition 1.6.1 (Strongly convex functions $\mathcal{SC}(\alpha)$) *We say that $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is α-strongly convex if $x \longrightarrow f(x) - \alpha\|x\|^2$ is convex.*

It is an easy exercice to check that α-strongly convex functions is equivalent to the following inequality :

$$f \in \mathcal{SC}(\alpha) \Longleftrightarrow \forall (x,y) \in \mathbb{R}^d \qquad f(x) \leq f(y) + \langle \nabla f(x), x - y\rangle - \frac{\alpha}{2}\|x - y\|^2.$$

This equivalence holds because of Proposition 1.4.2.

$$\begin{aligned}
f \in \mathcal{SC}(\alpha) &\iff f(.) - \frac{\alpha}{2}\|.\|^2 \text{ is convex} \\
&\iff \forall (x,y) \in \mathbb{R}^d \qquad f(y) - \frac{\alpha}{2}\|y\|^2 \geq f(x) - \frac{\alpha}{2}\|x\|^2 + \langle \nabla f(x) - \alpha x, y - x\rangle \\
&\iff \forall (x,y) \in \mathbb{R}^d \qquad f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}[\|y\|^2 - \|x\|^2] + \alpha[-\langle x, y\rangle + \|x\|^2] \\
&\iff \forall (x,y) \in \mathbb{R}^d \qquad f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|y - x\|^2
\end{aligned}$$

Another important criterion that implies strong convexity is concerned by the spectrum of the Hessian of f.

Proposition 1.6.1

$$f \in \mathcal{SC}(\alpha) \Longleftrightarrow \forall x \in \mathbb{R}^d \qquad D^2 f(x) \geqslant \alpha I_d,$$

where the last inequality holds w.r.t. the quadratic form inequalities.

Hence, f is α-strongly convex if and only if the spectrum of the Hessian of f is lower bounded by $\alpha > 0$ uniformly over $\mathbb{R}^d$.

Surrogates lower bounds As we already said in the previous paragraph, a smoothness property on the gradient function implies the existence of a surrogate function ϕ_+ :

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

Moreover, ϕ_- is also a surrogate function with $-$ instead of $+$ in front of $\frac{L}{2}\|x-y\|^2$. Now, if the function f is α strongly convex, then a second surrogate function $\tilde{\phi}_-$ also exists :

$$f(y) \geq \tilde{\phi}_-(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|x - y\|^2.$$

It is immediate to check that this last inequality with $\frac{\alpha}{2}$ is much more stronger than the one obtained with $-\frac{L}{2}$.

Another important property is given by the next result.

Proposition 1.6.2 *For any $f \in \mathcal{SC}(\alpha)$, we have*

$$\forall (x, y) \in \mathbb{R}^d \qquad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \alpha \|x - y\|^2 \tag{1.17}$$

Proof : The result is an easy consequence of

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

Indeed, if we switch x and y in the previous inequality, and add the two relationship, we obtain exactly the result we are looking for. □

1.6.2 Minimization of α-strongly convex and L-smooth functions

As we will see now, having both strong convexity and smoothness allows for a drastic improvement in the convergence rate. We denote $\kappa = \frac{\alpha}{L}$ for the condition number of f . The key observation is that a lower bound of $\langle \nabla f(x) - \nabla f(y), x - y \rangle$ can be improved and used in the proof of the convergence rate of the gradient descent.

We begin by the statement of the next Lemma.

Lemma 1.6.1 *Let f be a L-smooth and α-strongly convex function on $\mathbb{R}^d$, then*

$$\forall (x, y) \in \mathbb{R}^d \qquad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\alpha L}{\alpha + L}\|x - y\|^2 + \frac{1}{\alpha + L}\|\nabla f(x) - \nabla f(y)\|^2.$$

Proof : First step : auxiliary function φ. We consider $\varphi(x) = f(x) - \frac{\alpha}{2}\|x\|^2$ and apply Lemma 1.5.1 on φ, which is a convex function. First, remark that we necessarily have $L \geq \alpha$ because ∇f is L-Lipschitz and f lower bounded by $\alpha/2\|x\|^2$ when $x \longrightarrow +\infty$. Moreover, we have that φ is $L-\alpha$-smooth. Indeed, a direct computation shows that

$$\forall (x,y) \in \mathbb{R}^d \qquad 0 \leq \varphi(x) - \varphi(y) - \langle \nabla\varphi(y), x-y\rangle \leq \frac{L-\alpha}{2}\|x-y\|^2.$$

Symmetrizing in x and y, we get

$$\langle \nabla\varphi(x) - \nabla\varphi(y), x-y\rangle \leq (L-\alpha)\|x-y\|^2,$$

which shows that φ is $L-\alpha$-smooth.
Second step : coercivity of φ. Now, we can use Lemma 1.5.1 and obtain that

$$\forall (x,y) \in \mathbb{R}^d \qquad \varphi(x) - \varphi(y) \leq \langle \nabla\varphi(x), y-x\rangle - \frac{1}{2(L-\alpha)}\|\nabla\varphi(x) - \nabla\varphi(y)\|^2.$$

Again, symmetrizing in x and y and adding the two relationships, we obtain :

$$\langle \nabla\varphi(x) - \nabla\varphi(y), x-y\rangle \geq \frac{1}{L-\alpha}\|\nabla\varphi(x) - \nabla\varphi(y)\|^2. \tag{1.18}$$

Third step : algebraic conclusion. We replace now $\varphi(.)$ by its expression $f(.) - \frac{\alpha}{2}\|.\|^2$ and obtain

$$\begin{aligned}
(1.18) &\iff \langle \nabla f(x) - \nabla f(y) - \alpha(x-y), x-y\rangle \geq \frac{1}{L-\alpha}\|\nabla f(x) - \nabla f(y) - \alpha(x-y)\|^2 \\
&\iff \langle \nabla f(x) - \nabla f(y), x-y\rangle \left(1 + \frac{2\alpha}{L-\alpha}\right) \geq \left(\alpha + \frac{\alpha^2}{L-\alpha}\right)\|x-y\|^2 + \frac{\|\nabla f(x) - \nabla f(y)\|^2}{L-\alpha} \\
&\iff \langle \nabla f(x) - \nabla f(y), x-y\rangle \frac{L+\alpha}{L-\alpha} \geq \|x-y\|^2 \frac{L\alpha}{L-\alpha} + \frac{\|\nabla f(x) - \nabla f(y)\|^2}{L-\alpha} \\
&\iff \langle \nabla f(x) - \nabla f(y), x-y\rangle \geq \frac{\alpha L}{\alpha+L}\|x-y\|^2 + \frac{\|\nabla f(x) - \nabla f(y)\|^2}{L+\alpha}.
\end{aligned}$$

This last inequality is the desired conclusion. □

We then state the final result on the minimization of α-strongly convex function with a gradient descent algorithm 1.

Theorem 1.6.1 *Let f be a L-smooth and α-strongly convex function, then the choice of the step size $\gamma = \frac{2}{L+\alpha}$ leads to*

$$f(x_{n+1}) - f(x^\star) \leq \frac{L}{2}\exp\left(-\frac{4n}{\kappa+1}\right)\|x_1 - x^\star\|^2.$$

Proof : First, we shall remark that applying the L-smooth property given by Lemma 1.5.1 yields

$$f(x) - f(x^\star) \leq \frac{L}{2}\|x - x^\star\|^2$$

since $\nabla f(x^\star) = 0$.
We now use a recursion argument and write $\|x_{t+1} - x^\star\|^2$:

$$\begin{aligned}
\|x_{t+1} - x^\star\|^2 &= \|x_t - x^\star - \gamma\nabla f(x_t)\|^2 \\
&= \|x_t - x^\star\|^2 - 2\gamma\langle \nabla f(x_t), x_t - x^\star\rangle + \gamma^2\|\nabla f(x_t)\|^2 \\
&\leq \|x_t - x^\star\|^2 - 2\gamma\left[\frac{L\alpha}{L+\alpha}\|x_t - x^\star\|^2 + \frac{\|\nabla f(x_t)\|^2}{L+\alpha}\right] + \gamma^2\|\nabla f(x_t)\|^2,
\end{aligned}$$

where in the last line we applied Lemma 1.6.1. Then, we obtain

$$\|x_{t+1} - x^\star\|^2 \leq \left(1 - \frac{2\gamma L\alpha}{L+\alpha}\right)\|x_t - x^\star\|^2 + \|\nabla f(x_t)\|^2\left[\gamma^2 - \frac{2\gamma}{L+\alpha}\right]$$

Our choice $\gamma = \frac{2}{L+\alpha}$ permits to vanish the second term of the right hand side and we obtain :

$$\|x_{t+1} - x^\star\|^2 \leq \left(1 - \frac{2\gamma L\alpha}{L+\alpha}\right)\|x_t - x^\star\|^2 = \left(1 - \frac{2}{\kappa+1}\right)^2\|x_t - x^\star\|^2 \leq \exp\left(-\frac{4}{\kappa+1}\right)\|x_t - x^\star\|^2,$$

because $1 - x \leq e^{-x}$. Then, a simple recursion yields

$$\|x_{n+1} - x^\star\|^2 \leq \exp\left(-\frac{4n}{\kappa+1}\right)\|x_1 - x^\star\|^2$$

□

Computational complexity What should be kept in mind with this result is the strong improvement from a polynomial rate to an exponential one (the convergence is said to be linear) with the strongly convex property. Hence, in that case, with the gradient descent method, recovering an ϵ-solution of the minimization $\mathcal{P}$ requires $\frac{\kappa+1}{4}\log\epsilon^{-1}$ iterations instead of $L\epsilon^{-1}$ iterations in the simplest case of L-smooth function.

We should also remark that some lower bounds results exist on this type of class of convex functions. Indeed, even our results given in Theorem 1.5.1 and Theorem 1.6.1 are not optimal in the sense that some better algorithms exist for these classes of functions. In particular, it may be shown that a second order algorithm (called the Nesterov Accelerated Gradient Descent NAGD) may outperform the standard GD and attains the following rates of convergence :

— In the L-smooth case,

$$f(y_n) - f(x^\star) \leq \frac{2L}{n^2}\|y_1 - x^\star\|^2$$

— In the L-smooth and α-strongly convex situation :

$$f(y_n) - f(x^\star) \leq \frac{\alpha+L}{2}\|y_1 - x^\star\|^2\exp\left(-\frac{n-1}{\sqrt{\kappa}}\right)$$

Last but not least, it may be shown that this rates are optimal (see Nemirovski and Yudin, 1983) for these two classes of functions.

Chapitre 2

Stochastic optimization

In this chapter, we introduce the most important topic of this course : the stochastic optimization framework. We describe an important (though preliminary) result of almost sure convergence of stochastic gradient descent. We will obtain much more stronger results in the next chapter with the help of convex optimization.

2.1 Introductory example

2.1.1 Recursive computation of the empirical mean

The law of large number shows that estimating the mean of a distribution with the empirical mean is an efficient estimator... and it is also the first commonly used stochastic algorithm !

Consider a sequence of i.i.d. real variables (X_n) distributed according to μ, integrable whose expectation is denoted by$m$:

$$\mathbb{E}[X_1] = m.$$

The strong law of large number yields

$$\bar{X}_n := \frac{X_1 + \ldots + X_n}{n} \xrightarrow{n\to+\infty} m \quad a.s.$$

It is possible to re-write this sequence $(\bar{X}_n)_{n\in\mathbb{N}}$ with a recursive formulation :

$$\begin{aligned}\bar{X}_{n+1} &= \frac{n}{n+1}\bar{X}_n + \frac{1}{n+1}X_{n+1} \\ &= \bar{X}_n + \frac{1}{n+1}(X_{n+1} - \bar{X}_n).\end{aligned}$$

We can instantaneously remark that this sequence of empirical means is a Markov chain, where the innovation part is brought by the new observation X_{n+1} at time $n+1$.

Notation 2.1.1 (Filtration, step size) *The canonical filtration is denoted by $\mathcal{F}_n^X := \sigma(X_1, \ldots, X_n)$, and we define a sequence of step size*

$$\forall n \in \mathbb{N}^* \qquad \gamma_n := \gamma_1 n^{-\alpha}.$$

We consider h the function defined by :

$$\forall x \in \mathbb{R} \qquad h(x) = \frac{1}{2}\mathbb{E}[X_1 - x]^2.$$

Then, the sequence of empirical means may be written simply as :

$$\bar{X}_{n+1} = \bar{X}_n - \gamma_{n+1}\nabla_x h(\bar{X}_n) + \gamma_{n+1}\Delta M_{n+1},$$

where

$$\Delta M_{n+1} = -(X_{n+1} - \bar{X}_n) + \mathbb{E}[(X_{n+1} - \bar{X}_n)/\mathcal{F}_n] = -(X_{n+1} - \bar{X}_n) + \nabla_x h(\bar{X}_n).$$

Therefore, the element $\bar{X}_{n+1}$ is simply equals to $\bar{X}_n$ with the addition of a drift term (gradient descent of the function h) and a centered term, which will be considered as a martingale increment with respect to the filtration $(\mathcal{F}_n^X)_{n\geq 0}$

2.1.2 Recursive estimation of the mean and variance

We can also be interested in the simultaneous estimation of the mean m and the variance σ^2. We denote

$$S_n^2 := \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \bar{X}_n^2.$$

Then, we have

$$\begin{aligned}
S_{n+1}^2 &= \frac{1}{n+1}\left(\sum_{k=1}^{n}\left(X_k - \bar{X}_n - \frac{1}{n+1}[X_{n+1} - \bar{X}_n]\right)^2\right) \\
&= \frac{1}{n+1}\sum_{k=1}^{n}\left((X_k - \bar{X}_n)^2 + \frac{1}{(n+1)^2}[X_{n+1} - \bar{X}_n]^2 - \frac{2}{n+1}[X_k - \bar{X}_n][X_{n+1} - \bar{X}_n]\right) \\
&= \frac{n}{n+1}S_n^2 + \frac{1}{n+1}[X_{n+1} - \bar{X}_n]^2 + \frac{1}{(n+1)^2}[X_{n+1} - \bar{X}_n]^2 \\
&= S_n^2 - \gamma_{n+1}(S_n^2 - (\bar{X}_n - X_{n+1})^2) - \gamma_{n+1}^2(X_{n+1} - \bar{X}_n)^2.
\end{aligned}$$

If we denote now $Z_n = [\bar{X}_n, S_n^2]$, we obtain :

$$Z_{n+1} = Z_n - \gamma_{n+1}\left[H(Z_n, X_{n+1}) + (0, R_{n+1})\right]$$

where H is the function defined as :

$$H((\bar{x}, s^2), u) = (\bar{x} - u, s^2 - (\bar{x} - u)^2) \quad \text{et} \quad R_{n+1} = \gamma_{n+1}(X_{n+1} - \bar{X}_n)^2 \quad \text{(reste)}.$$

Again, we shall check that the joint evolution of $Z_n = [\bar{X}_n, S_n^2]$ is a Markov chain that may be written as :

$$\mathbb{E}[H(Z_n, X_{n+1}) \quad | \quad \mathcal{F}_n] = h(Z_n),$$

where :

$$h(\bar{x}, s^2) = (\bar{x} - m, s^2 - (\bar{x} - m)^2 - \sigma^2).$$

Again, if we denote $\Delta M_{n+1} = H(Z_n, X_{n+1}) - h(Z_n)$, then we still obtain a 2-dimensional martingale increment and we obtain the following representation :

$$Z_{n+1} = Z_n - \gamma_{n+1}h(Z_n) - \gamma_{n+1}(\Delta M_{n+1} + (0, R_{n+1})).$$

The strong law of large numbers permit to write the almost sure convergence :

$$Z_n \xrightarrow{n\to+\infty} (m, \sigma^2) = y^*$$

where y^* is the unique zero of the function h. In other words, this algorithm may be seen as a random algorithm for the computation of the zero of the function h.

2.1.3 Generic model of stochastic algorithm

The two algorithms above are typical examples of stochastic algorithms. The first one may be translated in the minimization of a function h while the second is concerned by the computation of the solutions of $h = 0$.

Definition 2.1.1 (Stochastic algorithm) *A stochastic algorithm is defined by :*

$$X_{n+1} = X_n - \gamma_{n+1} h(X_n) + \gamma_{n+1}(\Delta M_{n+1} + R_{n+1})$$

where (γ_n) *is a sequence of non negative step size such that*

$$\gamma_n \xrightarrow{n\to+\infty} 0 \quad \text{and} \quad \sum_{n\geq 1} \gamma_n = +\infty.$$

The sequence $(\Delta M_n)_{n\geq 0}$ *is a sequence of martingale increments and* $(R_n)_{n\geq 0}$ *is a sequence of perturbations (some neglictible rests) when* $n \to +\infty$.

The goal of this chapter is to describe the behaviour of such algorithms when the number of observations n becomes larger and larger. In particular, we will be interested in the following questions :

— convergence of the algorithm ?

— rate of convergence (with a central limit theorem) ?

At a later stage, we will also be interested in some non asymptotic results (with a finite horizon in n).

Concerning now the first question we will address in this chapter, a first heuristic answer is : if h has a repelling effect towards its minimum $x^\star$, then we can expect that

$$X_n \xrightarrow{n\to+\infty} x^\star,$$

if some technical conditions on $(\gamma_n)_{n\geq 1}$ are satisfied.

Concerning now the second question, we shall remark that in the two examples above, the almost sure convergence is given by the law of large number, so that the "rate" of convergence is $\sqrt{n}$. Again, this rate may certainly be related to the sequence of step size $(\gamma_n)_{n\geq 1}$.

2.2 Link with differential equation

If we consider the recursion :

$$x_{n+1} = x_n - \gamma_{n+1} h(x_n), \tag{2.1}$$

then this equation may be seen as an explicit Euler scheme that discretizes the following ordinary differential equation :

$$\frac{dy(t)}{dt} = -h(y_t). \tag{2.2}$$

In the case $\gamma_n = \gamma > 0$ and $h = -\nabla f$, we recover the situation illustrated in Chapter 1 with a gradient descent with a fixed step size.

In (2.1)-(2.2), we should understand the closeness of these two equations with an interpolation of the continous time evolution with a sequence of intervals whose size become smaller and smaller. If we define

$$\tau_n = \sum_{j=1}^{n} \gamma_j,$$

then we have the heuristic approximation $x_n = y(\tau_n)$ because when the step size are small enough (or equivalently when n is large enough) :

$$x_{n+1} = y(\tau_{n+1}) = y(\tau_n + \gamma_{n+1}) \simeq y(\tau_n) + \gamma_{n+1} y'(\tau_n) = x_n - \gamma_{n+1} h(y(\tau_n)) \simeq x_n - \gamma_{n+1} h(x_n).$$

Therefore, we can split the time interval as

$$[0, \tau_n] = [0, \tau_1] \cup [\tau_1, \tau_2] \ldots \cup [\tau_{n-1}, \tau_n].$$

For an infinite sequence of step size $(\gamma_n)_{n\geq 0}$, we can expect the evolutions of (2.1) and (2.2) under the *sine qua none* condition

$$\tau_n \to +\infty \qquad \text{meaning that} \qquad \sum_{j\geq 1} \gamma_j = +\infty.$$

2.3 Stochastic scheme

2.3.1 Motivations

The motivation for the study of the following evolution

$$X_{n+1} = X_n - \gamma_{n+1} h(X_n) + \gamma_{n+1}(\Delta M_{n+1} + R_{n+1}), \tag{2.3}$$

can be found in many practical situations.

Noisy gradients Let us imagine that h is the gradient of a function U we want to minimize. Without any loss of generality, we can assume that the minimizer is zero, so that the minimization of U and U^2 are equivalent. Now, we can imagine that when we are at step n at position X_n, the gradient of U is only accessible through unbiased measurement, then it is no longer possible to use the approach studied in Chapter 1. In such a case, the scheme is reduced to

$$X_{n+1} = X_n - \gamma_{n+1} h(X_n) + \gamma_{n+1} \xi_n,$$

which is closer to the formulation (2.3) with a null rest than the deterministic formulation (2.1).

Intractibility of the drift computation Let us imagine now that U is given through an integral :

$$\forall x \in \mathbb{R}^d \qquad U(x) = \int_E \mathcal{U}(x, y)\mu(dy),$$

where μ is a probability measure over E. The set is *a priori* known but possibly large so that it is difficult / impossible to conceive a numerical code that may compute the function h :

$$h(x) = \nabla U(x) = \int_E \partial_x \mathcal{U}(x, y)\mu(dy),$$

at every point $(X_n)_{n\geq 1}$ of the algorithm, because the computation over E is costly without any explicit formula for h.

In such a case, we can turn back to the formulation given by stochastic algorithm and remark that if $(Y_n)_{n\geq 1}$ is a sequence of i.i.d. random variables distributed according to μ over E, then the algorithm

$$X_{n+1} = X_n - \gamma_{n+1} \partial_x \mathcal{U}(X_n, Y_{n+1}),$$

may be written as

$$X_{n+1} = X_n - \gamma_{n+1} h(X_n) - \gamma_{n+1} \Delta M_{n+1},$$

because

$$\Delta M_{n+1} = h(X_n) - \partial_x \mathcal{U}(X_n, Y_{n+1}) = \int_E \partial_x \mathcal{U}(X_n, y)\mu(dy) - \partial_x \mathcal{U}(X_n, Y_{n+1})$$

is a martingale increment :

$$\mathbb{E}\left[\Delta M_{n+1} | \mathcal{F}_n\right] = \int_E \partial_x \mathcal{U}(X_n, y)\mu(dy) - \mathbb{E}\left[\partial_x \mathcal{U}(X_n, Y_{n+1}) | \mathcal{F}_n\right] = 0.$$

Conclusion : if we assume th we can sample μ, then if we denote by $(Y_n)_{n\geq 1}$ a sequence of i.i.d. random variables distributed according to μ, then we can define the sequence :

$$X_{n+1} = X_n - \gamma_{n+1} \frac{\partial \mathcal{U}}{\partial x}(X_n, Y_{n+1}),$$

which may re-written as

$$X_{n+1} = X_n - \gamma_{n+1} \nabla U(X_n) + \gamma_{n+1} \Delta M_{n+1}$$

whit

$$\Delta M_n = \nabla U(X_n) - \frac{\partial \mathcal{U}}{\partial x}(X_n, Y_{n+1}).$$

We recover a standard form of stochastic algorithm, which will be called « stochastic gradient descent ». We are interested in :

— Does (X_n) converges towards $x^\star$?

— What is the rate of convergence ? Under what type of conditions ? ... ?

2.3.2 Brief remainders on martingales

A good reference for a complete survey on martingales :
[W91] D. Williams *Probability with Martingales*, Cambridge Mathematical Textbooks, 1991.
You will find below a *pot-pourri* of important facts on martingales : some definitions, some important inequalities and some fundamental convergence theorems..

Definitions

Definition 2.3.1 (Martingale, Super-martingale, Sub-martingale) *A sequence $(X_n)_{n\geq 0}$ with values in $\mathbb{R}^d$ is a $(\mathcal{F}_n)$-martingale if*
(i) (X_n) is $(\mathcal{F}_n)$-measurable.
(ii) $\mathbb{E}[|X_n|] < +\infty$ for all $n \geq 0$.
(iii) $\mathbb{E}[X_{n+1}|\mathcal{F}_n] = X_n$ for all $n \geq 0$.

In the meantime, a real sequence $(X_n)_{n\geq 0}$ is a $(\mathcal{F}_n)$ super-martingale when (i), (ii) and (iii') are satisfied :

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] \leq X_n \qquad \forall n \geq 0.$$

Lastly, a real sequence $(X_n)_{n\geq 0}$ is a $(\mathcal{F}_n)$ sub-martingale when (i), (ii) and (iii") are satisfied

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] \geq X_n \qquad \forall n \geq 0.$$

Definition 2.3.2 (Predictible sequence) *We will say that a sequence $(Y_n)_{n\geq 0}$ is $(\mathcal{F}_n)_{n\geq 0}$ predictible if Y_{n+1} is $\mathcal{F}_n$-measurable, for all integer n :*

$$\mathbb{E}\left[Y_{n+1}|\mathcal{F}_n\right] = Y_{n+1}.$$

Proposition 2.3.1 (Doob decomposition) *Let (X_n) be a real sub-martingale (resp. real super-martingale). Then, a unique predictible increasing process (resp. decreasing process) denoted by $(A_n)_{n\geq 0}$ such that $A_0 = 0$ and $X_n = M_n + A_n$ where (M_n) is a $\mathcal{F}_n$-martingale. Moreover,*

$$A_n - A_{n-1} = \mathbb{E}[X_n - X_{n-1}|\mathcal{F}_{n-1}].$$

Definition 2.3.3 (Bracket of a square integrable martingale) *Let (M_n) be a real martingale that is square integrable. Then, (M_n^2) is a sub-martingale (because $x \mapsto x^2$ is convex). We denote by $(\langle M\rangle_n)$ the bracket of M, i.e. the unique predictible process vanishing at 0 such that $M_n^2 - \langle M\rangle_n$ is a martingale. We have :*

$$< M >_{n+1} - < M >_n = \mathbb{E}[(M_{n+1} - M_n)^2/\mathcal{F}_n].$$

Classical inequalities

Theorem 2.3.1 (Doob's inequalities) *We can state the following results, unavoidable in martingale's theory.*

Doob's inequality Let be given (X_n) a martingale or a sub-martingale. Then, we have for all $N \geq 0$, for all $p \geq 1$,

$$\mathbb{P}\left(\sup_{0\leq n\leq N} |X_n| \geq a\right) \leq \frac{1}{a^p}\mathbb{E}\left[|X_N|^p\right].$$

Doob's inequality in L^p Let be given (X_n) a martingale or a sub-martingale.

$$\mathbb{E}\left[\sup_{0\leq n\leq N} |X_n|^p\right] \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}[|X_N|^p]$$

In particular, if (X_n) is martingale vanishing at 0,

$$\mathbb{E}\left[\sup_{0\leq n\leq N} |X_n|^2\right] \leq \mathbb{E}[< X >_n^2]$$

and

$$\sup X_n \in L^p \text{ if and only if } \sup_{n\geq 0} \mathbb{E}[|X_n|^p] < +\infty.$$

Convergence results

Theorem 2.3.2 (Sub-martingale) *Let be given (X_n) a super-martingale or a sub-martingale such that*

$$\sup_{n\geq 1} \mathbb{E}[|X_n|] < +\infty.$$

Then, $(X_n)_{n\geq 0}$ converges a.s. towards an integrable random variable X_∞.

Theorem 2.3.3 (Super-Martingale) *We have the two fundamental results :*

i) Let (X_n) be a non-negative (super)-martingale, then (X_n) converges a.s.

ii) Moreover, assume that (X_n) is bounded in L^p with $p > 1$, then (X_n) converges in L^p.

Theorem 2.3.4 *Assume that (X_n) is a square integrable martingale. Then,*

i) On the event $\{< X >_\infty\} < +\infty\}$, (X_n) converges a.s. in $\mathbb{R}$.

ii) On $\{< X >_\infty = +\infty\}$, $\frac{X_n}{<X>_n} \xrightarrow{n\to+\infty} 0 \quad a.s.$

2.3.3 Robbins-Siegmund Theorem

We shall establish one of the most important result on stochastic algorithms. This result is known as the Robbins-Siegmund Theorem. We consider a measured space $(\Omega, \mathcal{F}, \mathbb{P})$.

Theorem 2.3.5 *(Robbins-Siegmund) Consider a filtration $(\mathcal{F}_n)_{n\geq 0}$ and four sequences of random variables (U_n), (V_n), (α_n) and (β_n) that are $(\mathcal{F}_n)$-measurables, non-negatives and integrables such that*

(i) (α_n), (U_n) and (β_n) are $(\mathcal{F}_n)$ predictables.

(ii) $\sup_{\omega\in\Omega} \prod_{n\geq 1} (1+\alpha_n(\omega)) < +\infty$ and $\sum_{n\geq 0} \mathbb{E}[\beta_n] < +\infty$.

(iii) $\forall n \in \mathbb{N}$,

$$\mathbb{E}[V_{n+1}|\mathcal{F}_n] \leq V_n(1+\alpha_{n+1}) + \beta_{n+1} - U_{n+1}.$$

Then,

$$\begin{cases} (a) & V_n \xrightarrow{n\to+\infty} V_\infty \in L^1 \quad \text{and} \quad \sup_{n\geq 0} \mathbb{E}[V_n] < +\infty. \\ (b) & \sum_{n\geq 0} \mathbb{E}[U_n] < +\infty \quad \text{and} \quad \sum_{n\geq 0} U_n < +\infty \quad a.s. \end{cases}$$

Remark 2.3.1 *Later on, the sequence (V_n) will be oftenly $V(X_n)$ where V is a "Lyapunov" function and (X_n) the state of the stochastic algorithm at step n. This result means that under Assumption (iii), we then deduce the almost sure convergence of $(V(X_n))$ when $n \to +\infty$. Therefore, under good assumptions on V, we then obtain the convergence of the algorithm itself (X_n).*

Proof : The idea behind the proof is to build a super-martingale from Inequality *(iii)*. This is in-line with the construction of a decreasing sequence (in the deterministic case).

Preliminary upper bound We shall remark that since $(U_n)_{n\geq 0}$ is predictable, then

$$\begin{aligned} \mathbb{E}\left[V_{n+1} + \sum_{k=1}^{n+1} U_k | \mathcal{F}_n\right] &\leq V_n(1+\alpha_{n+1}) + \sum_{k=1}^{n} U_k + \beta_{n+1}. \\ &\leq \left(V_n + \sum_{k=1}^{n} U_k\right)(1+\alpha_{n+1}) + \beta_{n+1}. \end{aligned}$$

Hence, if we define

$$S_n := \frac{V_n + \sum_{k=1}^n U_k}{\prod_{k=1}^n (1+\alpha_k)},$$

then we instantaneously remark that

$$\mathbb{E}[S_{n+1}|\mathcal{F}_n] \leq S_n + \tilde{\beta}_{n+1} \tag{2.4}$$

where

$$\tilde{\beta}_n = \frac{\beta_n}{\prod_{k=1}^n (1+\alpha_k)}.$$

Now, define $B_n := \sum_{k=1}^n \tilde{\beta}_k$, then the sequence $(B_n)_{n\geq 0}$ is a non-negative increasing sequence, which converges towards a random variable B_∞. Since $\tilde{\beta}_n \leq \beta_n$, assumption (ii) leads to $B_\infty \in L^1$:

$$\mathbb{E}B_\infty < +\infty.$$

We also obtain that from (2.4) that

$$\sup_{n\geq 0} \mathbb{E}[S_n] < +\infty.$$

Construction of a super-martingale We define $\tilde{S}_n = S_n + \mathbb{E}[B_\infty|\mathcal{F}_n] - B_n$. Since $(\beta_n)_{n\geq 0}$ is predictible, and because we have shown that $\mathbb{E}[S_{n+1}|\mathcal{F}_n] \leq S_n + \tilde{\beta}_{n+1}$, then we can write

$$\begin{aligned}\mathbb{E}[\tilde{S}_{n+1}|\mathcal{F}_n] &\leq S_n + \tilde{\beta}_{n+1} + \mathbb{E}[B_\infty|\mathcal{F}_n] - \mathbb{E}[B_{n+1}|\mathcal{F}_n] \\ &\leq S_n + \mathbb{E}[B_\infty|\mathcal{F}_n] - B_{n+1} + \tilde{\beta}_{n+1} = \tilde{S}_n.\end{aligned}$$

Moreover, we can check that $|\tilde{S}_n| \leq |S_n| + |\mathbb{E}[B_\infty|\mathcal{F}_n] - B_n|$ so that

$$\mathbb{E}|\tilde{S}_n| \leq \mathbb{E}S_n + \mathbb{E}B_\infty < +\infty,$$

We shall conclude that $(\tilde{S}_n)_{n\geq 0}$ is a super-martingale. Moreover, $(\tilde{S}_n)$ is clearly non-negative. Then, we can apply the convergence theorem on non-negative super-martingales. We then obtain :

$$\tilde{S}_n \xrightarrow{n\to+\infty} \tilde{S}_\infty \in L^1.$$

Moreover, since $\mathbb{E}[B_\infty|\mathcal{F}_n] - B_n \xrightarrow{n\to+\infty} 0$ a.s. and in L^1 (left as an exercice), we deduce that (S_n) converges a.s. towards $S_\infty := \tilde{S}_\infty$ with $S_\infty \in L^1$. Since $S_n \leq \tilde{S}_n$, we then conclude that

$$\sup_{n\geq 1} \mathbb{E}[S_n] < \mathbb{E}[\tilde{S}_1] < +\infty.$$

Going back to $(V_n)_{n\geq 0}$ and $(U_n)_{n\geq 0}$ We now focus our attention on the two sequences $(V_n)_{n\geq 0}$ and $(U_n)_{n\geq 0}$. The definition of S_n leads to

$$\mathbb{E}[V_n] + \mathbb{E}\left[\sum_{k=1}^n U_k\right] = \mathbb{E}[\left(\prod_{k=1}^n (1+\alpha_k)\right) S_n] \leq \left\|\prod_{k=1}^{+\infty}(1+\alpha_k)\right\|_\infty \mathbb{E}[S_n].$$

Then, we have shown that

$$\sup_{n\geq 1} \mathbb{E}[V_n] < +\infty \quad \text{and} \quad \mathbb{E}\left[\sum_{k\geq 1} U_k\right] < +\infty.$$

In particular, $\sum_{k\geq 1} U_k < +\infty$ *a.s.*, and we then obtain (b).

Next, since $S_n \to S_\infty$ and $\prod_{k\geq 1}(1+\alpha_k) < +\infty$, it easily follows that

$$S_n \prod_{k\geq 1}^n (1+\alpha_k) - \sum_{k\geq 1}^n U_k = V_n \xrightarrow{n\to+\infty} V_\infty = S_\infty \prod_{k\geq 1}(1+\alpha_k) - \sum_{k\geq 1} U_k$$

a.s. and in L^1. It proves (a) and concludes the proof. □

2.3.4 Application to stochastic algorithms

We now turn back to the initial motivation of stochastic recursive algorithms and consider the method defined by(2.3) with $X_0 = x_0 \in \mathbb{R}^d$ and

$$X_{n+1} = X_n - \gamma_{n+1} h(X_n) + \gamma_{n+1}(\Delta M_{n+1} + R_{n+1}) \tag{2.5}$$

where $(\gamma_n)_{n\geq 0}$ is a positive step-size sequence such that

$$\gamma_n \xrightarrow{n\to+\infty} 0, \quad \sum_{n\geq 1} \gamma_n = +\infty, \quad \sum_{n\geq 1} \gamma_n^2 < +\infty. \tag{2.6}$$

We will apply the Robbins-Siegmund convergence result to obtain the most important result of the chapter :

Theorem 2.3.6 (Convergence of the stochastic gradient method) *Let $(X_n)_{n\geq 0}$ be a sequence of random variables defined by (2.5), such that $(\gamma_n)_{n\geq 0}$ satisfies (2.6). Let V be a $\mathcal{C}^2$ L-smooth (sub-quadratic) function such that :*
$(\mathbf{H_1})$: *(« drift »assumption)*

$$m := \min_{x\in\mathbb{R}^d} V(x) > 0 \qquad \lim_{|x|\to+\infty} V(x) = +\infty, \quad \langle \nabla V, h\rangle \geq 0 \quad \text{and} \quad |h|^2 + |\nabla V|^2 \leq C(1+V).$$

$(\mathbf{H_2})$: *(Perturbations)*

(i) (ΔM_n) is a sequence of increments of an $(\mathcal{F}_n)$-martingale such that

$$\mathbb{E}[|\Delta M_n|^2|\mathcal{F}_{n-1}] \leq C(1+V(X_{n-1})) \qquad \forall n \in \mathbb{N}.$$

(ii) (R_n) is $(\mathcal{F}_n)$-measurable and

$$\mathbb{E}[|R_n|^2|\mathcal{F}_{n-1}] \leq C\gamma_n^2(1+V(X_{n-1})).$$

Then, under $(\mathbf{H_1})$ and $(\mathbf{H_2})$, then one has

$$(a)\sup_{n\geq 0}\mathbb{E}[V(X_n)] < +\infty, \quad (b)\sum_{n\geq 0}\gamma_{n+1}\langle\nabla V, h\rangle(X_n) < +\infty \quad a.s.$$
$$(c) V(X_n) \xrightarrow{n\to+\infty} V_\infty \in L^1 \quad a.s.,$$

$$(d) \qquad X_n - X_{n-1} \xrightarrow{n\to+\infty} 0 \quad a.s. \text{ and in } L^2.$$

Remark 2.3.2 *We shall make the following remarks.*

— *We can assume that X_0 itself is a random variable as soon as $\mathbb{E}[V(X_0)] < +\infty$.*

— *If the algorithm belongs to a convex set of $\mathbb{R}^d$, it is only needed that the assumptions on V are satisfied on this convex set.*

— *The consequence of*

$$\sum_{n\geq 0}\gamma_{n+1}\langle\nabla V, h\rangle(X_n) < +\infty \quad a.s.,$$

and $\langle\nabla V, h\rangle \geq 0$ and $\sum\gamma_n = +\infty$ is only

$$\liminf_{n\to+\infty}\langle\nabla V, h\rangle(X_n) = 0 \quad a.s.$$

But something more is needed to obtain the a.s. convergence of $\nabla V(X_n)$ towards 0.

Proof :

Below, we will use the following notation :

$$D^2V(x)y^{\otimes 2} = \sum_{k,l}\frac{\partial^2 V}{\partial x_k \partial x_l}(x)y_k y_l.$$

Moreover, C will denote a non explicit constant that may change from line to line.

Upper bound with the Taylor formula The second order Taylor formula shows the existence of a sequence $(\xi_n)_{n\geq 0}$ such that

$$V(X_{n+1}) = V(X_n) - \gamma_{n+1}\langle \nabla V(X_n), h(X_n)\rangle(X_n) + \gamma_{n+1}\langle \nabla V, \Delta M_{n+1}\rangle \\ + \gamma_{n+1}\langle \nabla V(X_n), R_{n+1}\rangle + \frac{1}{2}D^2V(\xi_{n+1})(\Delta X_{n+1})^{\otimes 2}$$

where ξ_{n+1} belongs to the segment $[X_n, X_{n+1}]$ and $\Delta X_{n+1} = X_{n+1} - X_n$. Since ∇V is L-Lipschitz, we know that D^2V is bounded (whatever the norm is) so that

$$\left|\frac{1}{2}D^2V(\xi_{n+1})(\Delta X_{n+1})^{\otimes 2}\right| \leq C|\Delta X_{n+1}|^2 \leq C_L\gamma_{n+1}^2(|h(X_n)|^2 + |\Delta M_{n+1}|^2 + |R_{n+1}|^2).$$

The assumption of the Robbins-Monro Theorem then shows that a large enough constant C exist such that :

$$\mathbb{E}\left[\left|\frac{1}{2}D^2V(\xi_{n+1})(\Delta X_{n+1})^{\otimes 2}\right| |\mathcal{F}_n\right] \leq C\gamma_{n+1}^2(1 + V(X_n)). \tag{2.7}$$

Finally, since ΔM_n is a martingale increment and V is lower bounded by $m > 0$, we can find a large constant C such that

$$\mathbb{E}\left[V(X_{n+1})|\mathcal{F}_n\right] \leq V(X_n)(1 + C\gamma_{n+1}^2) + \gamma_{n+1}\mathbb{E}\left[|\langle \nabla V(X_n), R_{n+1}\rangle||\mathcal{F}_n\right] - \gamma_{n+1}\langle \nabla V(X_n), h(X_n)\rangle. \tag{2.8}$$

From the Cauchy-Schwarz inequality and the assumption on the rest term R_n, we have :

$$\begin{aligned}
\mathbb{E}[|\langle \nabla V(X_n), R_{n+1}\rangle||\mathcal{F}_n] &\leq \mathbb{E}[|\nabla V(X_n)| \times |R_{n+1}||\mathcal{F}_n] \\
&= |\nabla V(X_n)|\mathbb{E}[|R_{n+1}||\mathcal{F}_n] \\
&\leq |\nabla V(X_n)|\sqrt{\mathbb{E}[|R_{n+1}|^2|\mathcal{F}_n]} \\
&\leq C\gamma_{n+1}\sqrt{1 + V(X_n)}|\nabla V(X_n)|.
\end{aligned}$$

Again, the function V being sub-quadratic, we have

$$|\nabla V(x)| = \mathcal{O}_{|x|\mapsto+\infty}(\sqrt{V(x)}),$$

so that a large enough constant C exists such that :

$$\mathbb{E}[|\langle \nabla V(X_n), R_{n+1}\rangle||\mathcal{F}_n] \leq C\gamma_{n+1}V(X_n).$$

Consequently, (2.8) implies that

$$\mathbb{E}\left[V(X_{n+1})|\mathcal{F}_n\right] \leq V(X_n)(1 + C\gamma_{n+1}^2) - \gamma_{n+1}\langle \nabla V(X_n), h(X_n)\rangle \tag{2.9}$$

Application of the Robbins-Siegmund Theorem We define the following quantities :

$$V_n := V(X_n), \quad U_{n+1} := \gamma_{n+1}\langle \nabla V, h\rangle(X_n), \quad \alpha_n := C\gamma_n^2, \quad \beta_{n+1} = 0,$$

and write (2.9) through the formulation (iii) of the Robbins-Siegmund Theorem :

$$\mathbb{E}[V_{n+1}|\mathcal{F}_n] \leq V_n(1 + \alpha_{n+1}) + \beta_{n+1} - U_{n+1}.$$

We can check rapidly that the needed assumptions are satisfied :

— $(\alpha_n)_{n\in\mathbb{N}}, (\beta_n)_{n\in\mathbb{N}}$ et $(U_n)_{n\in\mathbb{N}}$ are $(\mathcal{F}_n)_{n\in\mathbb{N}}$ predictible and (i) holds.

— The infinite product $\prod_{k=1}^{\infty}(1+\alpha_k)$ converges, leading to (ii).

Then, the Robbins-Siegmund Theorem implies that V_n converges towards V_∞ in L_1 and the series of U_n is a.s. convergent. We can then conclude that points (a), (b) and (c) hold. Concerning now (d), the arguments used above show that

$$\mathbb{E}[|\Delta X_{n+1}|^2] \leq C\gamma_{n+1}^2(1+\mathbb{E}[V(X_n)].$$

Next, the conclusion of (a) implies that $\sum \mathbb{E}[|\Delta X_k|^2]$ is a convergent series. We then conclude that

$$\mathbb{E}[|\Delta X_n|^2] \xrightarrow{n\to+\infty} 0 \quad \text{et} \quad \sum |\Delta X_{n+1}|^2 < +\infty \quad a.s.$$

In particular, $\Delta X_n \xrightarrow{n\to+\infty} 0$ a.s. (and in L^2). □

2.3.5 Unique minimizer

Corollary 2.3.1 *[Robbins-Monro Theorem] We assume that the assumptions of Theorem* (2.3.6) *hold. Moreover, we assume that h is continuous and that :*

$$\{x : \langle \nabla V(x), h(x)\rangle = 0\} = \{x^\star\}.$$

Then,
(α) $x^\star$ is the unique minimizer of V.
(β) $X_n \xrightarrow{n\to+\infty} x^\star$ a.s. and $\langle \nabla V, h\rangle(X_n) \xrightarrow{n\to+\infty} 0$.
(γ) If $p>0$ and $\rho \in [0,1)$ exists such that $\psi_p(x) = |x|^p$ with and $\psi_p(x) \leq CV^\rho(x)$, then,

$$\mathbb{E}[(\psi_p(X_n - x^\star)] \xrightarrow{n\to+\infty} 0.$$

Proof :
Point (α) : we know that V admits a minimizer and on this minimizer, one has $\langle \nabla V(x), h(x)\rangle = 0$. Our assumption then implies that $\{\nabla V = 0\} = \{x^\star\}$.
Point (β) : The point (b) and (c) above shows sthat

$$\sum_n \gamma_{n+1}\langle \nabla V, h\rangle(X_n) < +\infty \, a.s. \quad \text{and} \quad V(X_n) \to V_\infty \, a.s.$$

Hence, a subset exists $\tilde{\Omega} \subset \Omega$ with probability one $\mathbb{P}(\tilde{\Omega}) = 1$ and such that the two inequalities above hold for any $\omega \in \tilde{\Omega}$. In particular, considering such an event ω, we know that $(X_n(\omega))_n$ is a bounded sequence. Considering a convergent subsequence $(X_{\varphi(n)}(\omega))_n$ and its limit $X_\infty(\omega)$, we have $(\langle \nabla V, h\rangle(X_n))_n$ converges towards 0, and the continuity of h leads to $\langle \nabla V, h\rangle(X_\infty(\omega)) = 0$.

We then deduce that $X_\infty(\omega) = x^\star$ and the only possible accumulation point of $X_n(\omega)$ is $x^\star$. Moreover, $V(X_{\varphi(n)}) \to V(x^\star)$ and $V_\infty = V(x^\star)$ *p.s.* We can also conclude that (X_n) converges towards $x^\star$.
Point (γ) : we will use an equi-integrability argument. We fix $M>0$. The Lebesgue dominated convergence theorem shows that :

$$\mathbb{E}[(\psi_p(X_n - x^\star)1_{\psi_p(X_n - x^\star)\leq M}] \xrightarrow{n\to+\infty} 0.$$

Moreover, the Hölder inequality yields

$$\begin{aligned}
\mathbb{E}[\psi_p(X_n - x^\star)1_{\psi_p(X_n-x^\star)>M}] &\leq \mathbb{E}[(\psi_p(X_n - x^\star))^{1/\rho}]^\rho \mathbb{P}(\psi_p(X_n - x^\star) > M)^{1-\rho} \\
&\leq \sup_{n\geq 1} \mathbb{E}[V(X_n)]\mathbb{P}(\psi_p(X_n - x^\star) > M)^{1-\rho} \\
&\leq C\mathbb{P}(\psi_p(X_n - x^\star) > M)^{1-\rho}.
\end{aligned}$$

Now, the Lebesgue theorem implies that

$$\limsup_{M\to+\infty} \mathbb{E}[\psi_p(X_n - x^\star)1_{\psi_p(X_n-x^\star)>M}] = 0$$

and we obtain the result. □

2.3.6 Isolated critical points

We now derive some results on the asymptotic behaviour of the Robbins-Monro algorithm when the set of minimizers (critical points indeed) is isolated and finite.

Corollary 2.3.2 *Under the assumptions of Theorem* (2.3.6)*, assume moreover that* $(\mathbf{H_{Finite}})$: *h is continuous and for all $v \geq 0$, $\{x, V(x) = v\} \cap \{\langle \nabla V, h\rangle = 0\}$ is finite. Then, (X_n) converges towards X^∞ a.s. and $\langle \nabla V, h\rangle(X_\infty) = 0$.*

Proof :

Topology of the adherence We know that $V(X_n) \xrightarrow{n\to+\infty} V_\infty$ a.s., which is a finite random variable and $\lim V(x) = +\infty$ when $|x| \to +\infty$. Therefore, we can find $\tilde{\Omega} \subset \Omega$ of probability 1 such that $\omega \in \tilde{\Omega}$ for which $(X_n(\omega))_{n\in\mathbb{N}}$ is a bounded sequence. We denote by χ^∞ the set of possible accumulation points (with a fixed ω). Then, this set is bounded and closed. Hence, χ^∞ is a compact set.

Moreover, we can restrict our study to the events ω such that $\Delta X_n \to 0$ because this convergence holds almost surely. This last point then implies that χ^∞ is a connected set. To show this last point, assume that $\chi^\infty$is not connected. It then implies that two non-empty and disjoint closed sets F_1 and F_2 exist such that $\chi^\infty = F_1 \cup F_2$. Consider $x \in F^1 \cap \chi^\infty$ and $y \in F^2 \cap \chi^\infty$. The definition of χ^∞ yields the existence of two sub-sequences $(X_{\phi_x(n)})$ and $(X_{\phi_y(n)})$ that converge respectively towards x and y. Since F_1 and F_2 are closed and disjoints, we know that $d(F_1, F_2) = d_0 > 0$. We also know that for all $\varepsilon > 0$, an integer $n_0 \in \mathbb{N}$ exists such that $n \geq n_0$,

$$|X_{\phi_x(n)} - x| \leq \varepsilon, \quad |X_{\phi_y(n)} - y| \leq \varepsilon \quad \text{and} \quad |X_n - X_{n-1}| \leq \varepsilon.$$

We consider for example $\varepsilon = d_0/4$ and consider the sequence of times (T_n^x) and (T_n^y) as follows :

$$\begin{aligned} T_1^x &:= \inf\{n \geq n_0, |X_n - x| \leq d_0/4\}, \quad T_1^y := \inf\{n \geq T_1^x, |X_n - y| \leq d_0/4\} \\ T_k^x &:= \inf\{n \geq T_{k-1}^y, |X_n - x| \leq d_0/4\}, \quad T_k^y := \inf\{n \geq T_k^x, |X_n - y| \leq d_0/4\}. \end{aligned}$$

We clearly have that T_k^x and T_k^y are finite for all k because x and y belong to χ^∞. Since $|X_n - X_{n-1}| \leq d_0/4$ for all $n \geq n_0$, we can deduce that in between T_k^x and T_k^y, an integer n_k exists such that $d(X_{n_k}, F_1 \cup F_2) > d_0/4$. Since the sequence (X_{n_k}) is bounded, it has an accumulation point X_∞. By construction, it is clear that X_∞ does not belong to $F_1 \cup F_2$, which is a contradiction.

Almost sure convergence We consider the trajectories such that $V(X_n) \to V_\infty$, we then deduce that $\chi^\infty(\omega) \subset \{x, V(x) = V_\infty(\omega)\}$. Since

$$\sum \gamma_k \langle \nabla V, h\rangle(X_k) < +\infty \quad a.s.,$$

we know that a point $y^\star$ of $\chi^\infty(\omega)$ exists such that

$$y^\star \in \{\langle \nabla V, h\rangle = 0\}.$$

It is much more complicate to show that every point of $\chi^\infty(\omega)$ satisfies this property. We will establish such a property only in dimension 1 (the case of higher dimensions is more involved and requires some ingredients related to pseudo-trajectories of differential systems).

We consider the two cases :

— If $\chi^\infty(\omega)$ is reduced to a singleton, then the proof is complete.

— Otherwise, $\chi^\infty(\omega)$ is a connected of $\mathbb{R}$ with a non-empty interior. In particular, since V is necessarily constant on $\chi^\infty(\omega)$ and V' is a continuous function, it implies that $V' = 0$ on $\chi^\infty(\omega)$. It then implies that $\langle \nabla V, h\rangle(x) = 0$ fpr all $x \in \chi^\infty(\omega)$ so that $\chi^\infty(\omega)$ is included in a connected component of $\{x, V(x) = V_\infty(\omega)\} \cap \{\langle \nabla V, h\rangle = 0\}$. Our assumption then implies that this set is indeed locally finite. Hence, $\chi^\infty(\omega)$ is solely reduced to a single point and the proof is complete.

□

Problem : When we are considering the Robbins-Monro algorithm, we have shown that the algorithm converges towards a critical point. However, it is not clear that such a convergence holds towards a local minimum of the function V. It is however possible to establish this kind of result with some extra-assumptions. This is beyond the scope of this courses but some ingredients may be found in the Lecture Notes of M. Benaïm (pseudo-trajectories, Kushner-Clark Theorem, local traps avoided by stochastic algorithms, ...)

Example 2.3.1 *Recursive least squares. We consider $\xi_1, \dots, \xi_n, \dots$ a set of observation vectors of $\mathbb{R}^d$ (the inputs) and we are looking for finding a relationship $\xi \to F(\xi)$ with a linear model that minimizes the least squares criterion, i.e. we want to minimize :*

$$V : C \in \mathbb{R}^d \mapsto \sum_{k=1}^{n} (F(\xi_k) - \langle C, \xi_k\rangle)^2 = \mathbb{E}_\mu \left[(F(\xi) - \langle C, \xi\rangle)^2 \right]$$

where $\mu = 1/n \sum_{k=1}^n \delta_{\xi_k}$. If we denote by $A(\xi) := \xi\xi^t$ the Gramm matrix, then some immediate computations show that :

$$\nabla V(C) = \mathbb{E}_\mu[(\langle C, \xi\rangle - F(\xi))\,\xi], \quad (D^2 V(C)) = \mathbb{E}_\mu[A(\xi)] = cte.$$

Moreover, D^2V is clearly symmetric and non-negative. To obtain its invertibility, we need that u belongs to the orthogonal of the set spanned by ξ_k. Therefore, V is strictly convex if and only if $(\xi_k)_{1\le k\le n}$ generates $\mathbb{R}^d$. In this case, V admits a unique minimizer. In such a case, we can use the following recursive formulation to compute (C_n) with

$$C_{n+1} = C_n - \gamma_{n+1}(\langle C_n, \xi_{n+1}\rangle - F(\xi_{n+1}))\xi_{n+1}$$

where (ξ_n) is an i.i.d. sequence of random variables distributed according to μ. We can check the assumptions of the Robbins-Monro Theorem and then conclude that C_n converges a.s. towards the unique minimizer of V.

We will be interested later in the **rates of convergence** possibly attained by such algorithms. Such rates will strongly rely on convex properties of the function we are interested in. In particular, we will derive some convergence rates for convex and strongly convex functions.

Chapitre 3

Non-asymptotic study of stochastic algorithms

3.1 Introduction