M2RI UT3
S10 - Stochastic Optimization algorithms

S. Gadat

Toulouse School of Economics

Université Toulouse I Capitole

4 février 2018

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 $(X_1, Y_1), \ldots, (X_n, Y_n)$.

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 $(X_1, Y_1), \ldots, (X_n, Y_n)$ 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 Optimization of the likelihood

1.1.1.1 Definitions and important properties

We introduce in this paragraph an instant crush on the usefulness of the likelihood in statistics. We consider a sequence of i.i.d. observations $X_1, \ldots, X_n$ with a distribution function $f(x, \theta_0)$. We want to estimate θ_0 the true value of the parameter in a set of possible values Θ. The joint density of a n-sample $(x_1, \ldots, x_n)$ is then

$$f_n(x_1, \ldots, x_n|\theta) = f(x_1, \theta) \times f(x_2, \theta) \ldots f(x_n, \theta).$$

The likelihood function L_n is defined as

$$L_n(\theta) = \prod_{i=1}^{n} f(X_i, \theta),$$

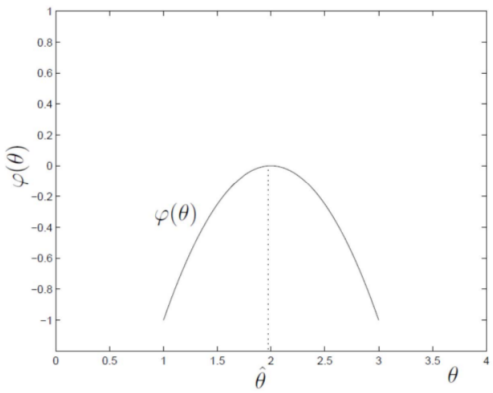

FIGURE 1.1: Typical behaviour of "good" likelihood.

while the log-likelihood function l_n is simply the logarithm of the previous function

$$l_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^{n} \log f(X_i, \theta).$$

The M.L.E. $\hat{\theta}$ is the value that maximizes the function $\theta \mapsto L_n(\theta)$. The typical situation is presented in Figure 1.1.

Said differently, the MLE estimator is defined as the value $\hat{\theta}$ in Θ that is the most likely to produce the set of n observations $(X_1, \ldots, X_n)$.

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L_n(\theta) = \arg\max_{\theta \in \Theta} l_n(\theta).$$

We present a brief pot-pourri of nice properties of the MLE below :

— **Consistency :** the MLE satisfies

$$\mathbb{P}_{\theta_0}\left[\lim_{n \longrightarrow +\infty} \hat{\theta}_n = \theta_0\right] = 1$$

— **Asymptotic normality :**

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \longrightarrow \mathcal{N}(0, \sigma^2_{MLE})$$

— **Asymptotic optimality :** in a generic settings, the MLE has the smallest possible variance of estimation in the class of all asymptotically unbiased estimators of θ_0. In a sense, there is nothing better to expect with another different statistical estimators. Such a sentence should indeed be balanced by several considerations : computational cost, regularity assumptions on $\theta \longmapsto f(x, \theta)$.

1.1.1.2 Why it works ?

Behaviour of the scaled log-likelihood The feeling behind the good behaviour of $\hat{\theta}$ is that the scaled log-likelihood converges towards a deterministic function when $n \longrightarrow +\infty$:

$$\frac{1}{n}\sum_{i=1}^{n} \log f(X_i, \theta) \longrightarrow \mathbb{E}_{\mathbb{P}_{\theta_0}}[\log f(X, \theta)],$$

with the help of the law of large numbers. It is important to understand that the expectation above has to be taken with respect to the distribution $\mathbb{P}_{\theta_0}$ since it is the underlying (unknown) distribution of the observations.

The important function is then the function defined by

$$\ell(\theta) := \int_{x \in X} \log(f(x,\theta) f(x,\theta_0) dx = \mathbb{E}_{\mathbb{P}_{\theta_0}}[\log f(X,\theta)],$$

which is the exact deterministic counterpart function of the scaled log-likelihood. In particular, it is possible to prove easily that θ_0, the true value of the hidden parameter, is the position that maximizes ℓ :

$$\begin{aligned} \forall \theta \in \Theta \qquad \ell(\theta) - \ell(\theta_0) &= \int_{x \in X} [\log f(x,\theta) - \log f(x,\theta_0)] f(x,\theta_0) dx \\ &= \int_{x \in X} \log \left[\frac{f(x,\theta)}{f(x,\theta_0)} \right] f(x,\theta_0) dx \\ &\leq \int_{x \in X} \left[\frac{f(x,\theta)}{f(x,\theta_0)} - 1 \right] f(x,\theta_0) dx, \end{aligned}$$

where we used the inequality $\log t \leq t - 1$. Then,

$$\forall \theta \in \Theta \qquad \ell(\theta) - \ell(\theta_0) \leq \int_{x \in X} [f(x,\theta) - f(x,\theta_0)] dx = 1 - 1 = 0.$$

Hence,

$$\theta_0 = \arg\max_{\theta \in \Theta} \ell(\theta).$$

Convergence Therefore, the idea is that something like what is illustrated in Figure 1.2 occurs.

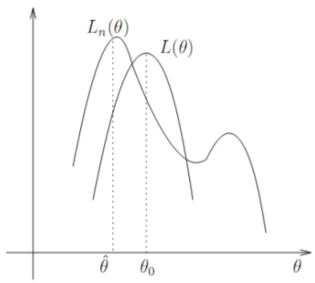

FIGURE 1.2: Convergence of the scaled log-likelihood (and of the likelihood itself) function when n increases.

A typical behaviour of the likelihood function when the number of observations increases is described in Figure 1.3. The likelihood function becomes more and more concentrated near the true value of the parameter θ_0.

Thus, a cornerstone of several problems in statistics will be how to write a statistical model with a good likelihood function and how to design efficient algorithms for solving the maximization problem associated to the definition of $\hat{\theta}$. In some cases, this maximization yields an explicit formula while in some other cases, this exact maximization is not solvable with a direct formula. In the next paragraphs, we provide two typical famous examples.

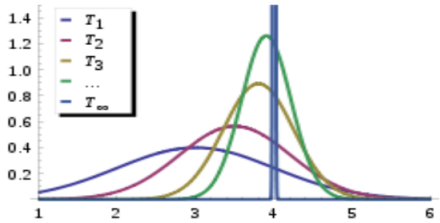

FIGURE 1.3: Typical behaviour of the variations of the likelihood function when n increases.

1.1.2 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 the maximizer of

$$U(\theta) = \|Y - X\theta\|_2^2,$$

where $\|.\|_2^2$ refers to the Euclidean L^2 norm of vectors in $\mathbb{R}^n$. In particular,

$$U(\theta) =^t (Y - X\theta)(Y - X\theta) =^t YY - 2^t YX\theta +^t \theta^t XX\theta.$$

It is possible to prove that U is a convex function (quadratic) and then maximizing U is then equivalent solving

$$DU(\theta) = 0.$$

Such an equation has an explicit solution because :

$$DU(\theta) = -2^t YX + 2(^t XX)\theta.$$

Hence, the MLE $\hat{\theta}$ is obtained with :

$$\hat{\theta}_n := ({}^tXX)^{-1}\,{}^tXY. \tag{1.2}$$

Remark 1.1.1 *We should remark that Equation 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 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.

1.1.3 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

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

1.1.3.1 Homogeneous case

In a first time, we assume that the probability of success of Y is independent of X, then $p(X, \theta_0) = p_0$ and we want to recover the value of p_0 from the observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. In that case, if we introduce $S_n = \sum_{i=1}^{n} Y_i$, then

$$\ell(\theta) = S_n \log p + (n - S_n) \log 1 - p.$$

The derivation of ℓ is easy :

$$\ell'(\theta) = \frac{S_n}{p} - \frac{n - S_n}{1 - p}$$

Solving $\ell'(\theta) = 0$ is possible :

$$\ell'(\theta) = 0 \Leftrightarrow (1 - p)S_n = p(n - S_n) \Leftrightarrow S_n = np \Leftrightarrow p = \frac{S_n}{n}$$

The M.L.E. in the classification model is then estimated by :

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

which is the mean number of success in the sample (rather obvious result !)

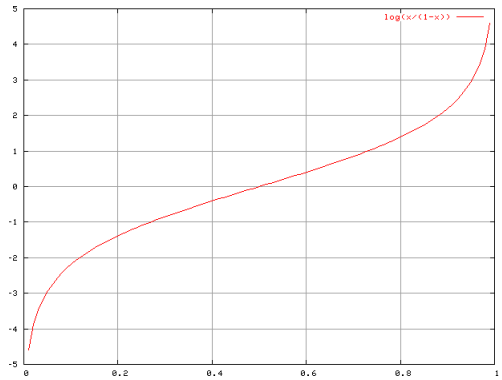

FIGURE 1.4: Baseline Logit function when p varies between 0 and 1.

1.1.3.2 Inhomogeneous case

Logit model 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 $[-\infty, \infty]$. This function is shown in Figure 1.4.

The logistic regression model is then defined by :

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

We expect a linear relationship between the logit(p) and the set of variables gathered in $x \in \mathbb{R}^p$.

An easy resolution leads to

$$\frac{p}{1-p} = e^{\langle \theta_0, x \rangle} \Longleftrightarrow p = e^{\langle \theta_0, x \rangle}(1-p) \Longleftrightarrow p(1 + e^{\langle \theta_0, x \rangle}) = e^{\langle \theta_0, x \rangle}.$$

Therefore, we obtain that

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

In the same time, we also prove that

$$\mathbb{P}(Y = 0|X = x) = \frac{1}{1 + e^{\langle \theta_0, x \rangle}}.$$

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.

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

$$L(\theta) = \prod_{i=1}^{n} \left[\frac{e^{\langle \theta, X_i \rangle}}{1 + e^{\langle \theta, X_i \rangle}}\right]^{Y_i} \left[\frac{1}{1 + e^{\langle \theta, X_i \rangle}}\right]^{(1-Y_i)} = \prod_{i=1}^{n} \left[\frac{e^{Y_i \langle \theta, X_i \rangle}}{1 + e^{\langle \theta, X_i \rangle}}\right].$$

Thus, the log-likelihood function is given by

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

The first order differentiate function of $\psi : \theta \longrightarrow \log(1 + e^{\langle \theta, X_i \rangle})$ is easy to compute :

$$\partial_{\theta_k} \psi(\theta) = X_{i,k} \frac{e^{\langle \theta, X_i \rangle}}{1 + e^{\langle \theta, X_i \rangle}}.$$

Unfortunately, it seems a little bit difficult to find all the solutions of $\partial_{\theta_k} \ell(\theta) = 0$ because it leads to a highly non-linear system of equations :

$$\forall k \in \{1, \ldots, p\} \qquad \partial_{\theta_k} \ell(\theta) = 0 \Longleftrightarrow \forall k \in \{1, \ldots, p\} \qquad \sum_{i=1}^{n} Y_i X_{i,k} - \sum_{i=1}^{n} X_{i,k} \frac{e^{\langle \theta, X_i \rangle}}{1 + e^{\langle \theta, X_i \rangle}} = 0.$$

However, it is easy to show that $\psi : \theta \longrightarrow \log(1 + e^{\langle \theta, X_i \rangle})$ is a convex function, so that $\theta \longrightarrow -\log(1 + e^{\langle \theta, X_i \rangle})$ is concave. We have already computed the first order differentiate function $D\psi = [\partial_{\theta_1} \psi, \ldots, \partial_{\theta_k} \psi]$. The second order differentiate function is

$$D^2 \psi = \left(\partial_{\theta_k} \partial_{\theta_l} \psi\right)_{k,l},$$

which is computed as

$$\forall (k,l) \in \{1, \ldots, p\}^2 \qquad \partial_{\theta_k} \partial_{\theta_l} \psi = X_{i,k} X_{i,l} \frac{e^{\langle \theta, X_i \rangle}}{1 + e^{\langle \theta, X_i \rangle}} - X_i X_j \frac{e^{\langle \theta, X_i \rangle} \times e^{\langle \theta, X_i \rangle}}{(1 + e^{\langle \theta, X_i \rangle})^2} = X_{i,k} X_{i,l} \frac{e^{\langle \theta, X_i \rangle}}{(1 + e^{\langle \theta, X_i \rangle})^2}.$$

In particular, we can prove that

$$\forall u \in \mathbb{R}^p \qquad {}^t u D^2 \psi u \geq 0,$$

which implies that ψ is convex. Such a property will be detailed later on during the Lecture Notes. 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.

1.1.4 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.

1.1.4.1 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 Equation (1.1) and solved by Equation (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...

1.1.4.2 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.

1.1.4.3 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.

<u>*Proof :*</u> 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. $\square$

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 $f(x) = x_1^2 + a x_1 x_2 + b x_1 + x_2^2$, then

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

1.3.2 Smoothness class and consequences

We define below the set of functions $\mathcal{C}_L^1$ 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 $f \in \mathcal{C}_L^1$, we have*

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

<u>*Proof :*</u> 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 :*

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

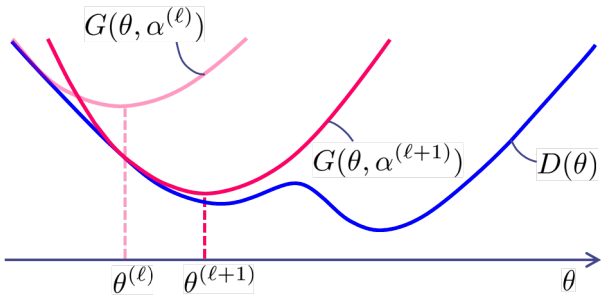

FIGURE 1.5: Geometrical illustration of the MM algorithm : we minimize $\theta \longrightarrow D(\theta)$ with the help of some auxiliary functions $\theta \longrightarrow G(\theta, \alpha)$.

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 1.5.

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

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

— 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 Equation (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) :

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

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

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

   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) :

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

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 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 Equation (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 Equation (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 Equation (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 Equation (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.

<u>Second step : $\|x_s - x^\star\|$ is a decreasing sequence.</u> 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.
<u>Third step : Convergence rate of $(\delta_s)_{s \geq 1}$.</u> We know from Equation (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 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.

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

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

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

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 Equation (2.1)-Equation (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 Equation (2.1) and Equation (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 Equation (2.3) with a null rest than the deterministic formulation Equation (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.*

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

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

<u>Doob's inequality in L^p</u> 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 $\{< X >_\infty = +\infty\}$, $\frac{X_n}{<X>_n} \xrightarrow{n\to+\infty} 0$ 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,

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

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 Equation (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 byEquation (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 Equation (2.5), such that $(\gamma_n)_{n\geq 0}$ satisfies Equation (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, Equation (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 Equation (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 Equation (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 Equation (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 :

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

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

3.1.1 Choice of the step size

In this chapter, we will specifically address the problem of estimating convergence rate of the recursive estimates derived from stochastic algorithms. In particular, we are interested in the non asymptotic behaviour of the stochastic gradient descent

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

This behaviour will highly depend on the step-size sequence chosen in the recursive formula. To understand this dependency, we consider again the formula given by the empirical mean :

$$\theta_0 = 0 \qquad \text{and} \qquad \theta_n = \theta_{n-1} - \gamma_n(\theta_{n-1} - Z_n). \tag{3.1}$$

If we choose $\gamma_n = \frac{1}{n}$, then we recover the standard average :

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

But we may choose other step-size sequences, according to what we have seen in the previous chapter. Choosing now $\gamma_n = \frac{2}{n+1}$, we can show using a recursion that

$$\theta_n = \frac{2}{n(n+1)}\sum_{k=1}^{n} kx_k.$$

Of course, the behaviour of this last algorithm may be different from the one using the step-size $\gamma_n = n^{-1}$. We will try to understand the typical behaviour of the stochastic algorithms with respect to this key choice.

3.1.2 Linear case

Since we plan to derive a good understanding of the situation with respect to the sequence $(\gamma_n)_{n\geq 1}$, we first study the simplest situation that corresponds to the *linear* case : $\nabla f(x) = x$.

Let us have a look at Equation Equation (3.1) while assuming that $(Z_i)_{i\geq 1}$ are all i.i.d. and $\mathbb{E}[Z] = m$ and with a variance σ^2. Here, we can directly study what happens in $\mathbb{R}^d$ instead of

$\mathbb{R}$ because of the very simple relationship between the spectral decomposition of $\nabla f(x)$ and $\mathbb{R}^d$! Note that it will not always be the case below. We can write

$$\begin{aligned}
\theta_n - m &= \theta_{n-1} - m - \gamma_n(\theta_{n-1} - Z_n) \\
&= (\theta_{n-1} - m)(1-\gamma_n) + \gamma_n(Z_n - m) \\
&= \dots \\
&= \prod_{k=1}^{n}(1-\gamma_k)(\theta_0 - m) + \sum_{i=1}^{n}\left[\prod_{k=i+1}^{n}(1-\gamma_k)\right]\gamma_i(Z_i - m).
\end{aligned}$$

We could understand the previous decomposition as a bias/variance decomposition. The first term only depends on the initialization of the algorithm, and is decreased according to the effect of the product $\prod_{k=1}^{n}(1-\gamma_k)$. The second term is a variance term, because it involves the observations $(Z_i)_{i\geq 1}$.

In particular, we can compute the quadratic risk of estimation :

$$\begin{aligned}
\mathbb{E}|\theta_n - m|^2 &= \mathbb{E}\left|\prod_{k=1}^{n}(1-\gamma_k)(\theta_0 - m) + \sum_{i=1}^{n}\left[\prod_{k=i+1}^{n}(1-\gamma_k)\right]\gamma_i(Z_i - m)\right|^2 \\
&= \mathbb{E}\left|\prod_{k=1}^{n}(1-\gamma_k)(\theta_0 - m)\right|^2 + 2\mathbb{E}\left|\prod_{k=1}^{n}(1-\gamma_k)(\theta_0 - m)\sum_{i=1}^{n}\left[\prod_{j=i+1}^{n}(1-\gamma_j)\right]\gamma_i(Z_i - m)\right| \\
&\quad + \mathbb{E}\left|\sum_{i=1}^{n}\left[\prod_{k=i+1}^{n}(1-\gamma_k)\right]\gamma_i(Z_i - m)\right|^2 \\
&= \prod_{k=1}^{n}(1-\gamma_k)^2(\theta_0 - m)^2 + \sum_{i=1}^{n}\gamma_i^2\prod_{k=i+1}^{n}(1-\gamma_k)^2\sigma^2,
\end{aligned}$$

where the last line comes from the fact that the sequence of observations $(Z_i)_{i\geq 1}$ are assumed to be i.i.d., centered at m, and with a variance σ^2 :

$$\mathbb{E}|Z - m|^2 = \sigma^2.$$

To understand now the behaviour of the quadratic risk, we need to understand the size of the bias and of the variance. Indeed, if we introduce the time of the algorithm defined by :

$$\Gamma_n := \sum_{k=1}^{n}\gamma_k,$$

we then obtain

$$\mathbb{B}_n = \prod_{k=1}^{n}(1-\gamma_k)^2(\theta_0 - m)^2 \leq (\theta_0 - m)^2\prod_{k=1}^{n}e^{-2\gamma_k} = (\theta_0 - m)^2 e^{-2\Gamma_n},$$

while

$$\mathbb{V}_n = \sigma^2\sum_{i=1}^{n}\gamma_i^2\prod_{k=i+1}^{n}(1-\gamma_k)^2 \leq \sigma^2\sum_{i=1}^{n}\gamma_i^2 e^{-2(\Gamma_n - \Gamma_i)}.$$

Bias term Now, a good behaviour of the bias term is certainly related to the increase of $n \longrightarrow \Gamma_n$. Note that the inequality above is indeed sharp since $1 - x \sim e^{-x}$ when $x \longrightarrow 0$.

We have seen that a good use of the Robbins-Monro algorithm was obtained by choosing $\sum \gamma_n = +\infty$. This is indeed recovered through our analysis of our bias term. Moreover, we can push the non-asymptotic analysis further and show the next proposition.

Proposition 3.1.1 *If we choose $\gamma_n = \gamma n^{-\alpha}$ with $\alpha \in (0, 1]$, then :*

— *If $\alpha < 1$, then*

$$\sum_{k=1}^{n} \gamma_k \geq \frac{\gamma}{1-\alpha} n^{1-\alpha} \qquad \textit{and} \qquad \mathbb{B}_n \leq |\theta_0 - m|^2 e^{-\frac{2\gamma}{1-\alpha} n^{1-\alpha}}$$

.

— *If $\alpha = 1$, then*

$$\sum_{k=1}^{n} \gamma_k \geq \gamma \log(n) \qquad \textit{and} \qquad \mathbb{B}_n \leq |\theta_0 - m|^2 e^{-2\gamma \log(n)}$$

.

Proof : We show this lower bounds using a series/integral comparison argument. We remark that $x \longrightarrow x^{-\alpha}$ is decreasing, so that

$$\sum_{k=1}^{n} \gamma_k = \gamma \sum_{k=1}^{n} k^{-\alpha} \geq \gamma \int_1^n x^{-\alpha} dx.$$

This ends the proof. □

Variance term The variance term may also be upper bounded following the same kind of strategy.

Proposition 3.1.2 *If we choose $\gamma_n = \gamma n^{-\alpha}$ with $\alpha \in (0, 1]$, then :*

— *If $\alpha < 1$, then an integer n_0 exists such that*

$$\forall n \geq n_0 \qquad \mathbb{V}_n \leq 2\sigma^2 \gamma_n$$

.

— *If $\alpha = 1$ and $\gamma \neq 1/2$, then*

$$\mathbb{V}_n \leq 2\sigma^2 \frac{\gamma^2}{n}$$

.

— *If $\alpha = 1$ and $\gamma = 1/2$, then*

$$\mathbb{V}_n \leq \sigma^2 \frac{\log n}{2n}$$

.

Proof : We compute an upper bound of $\mathbb{V}_n$ and we write with $\gamma_n = \gamma n^{-\alpha}$ if $\alpha < 1$:

$$\begin{aligned} \sum_{k=1}^{n} \gamma_k^2 \prod_{l=k+1}^{n} (1 - \gamma_l)^2 &\leq \sum_{k=1}^{n} \gamma_k^2 e^{-2\sum_{k+1}^{n} \gamma_l} \\ &= \sum_{k=1}^{n} \gamma_k^2 e^{-2\Gamma_n + 2\Gamma_k} \leq \gamma^2 e^{-2\Gamma_n} \sum_{k=1}^{n} k^{-2\alpha} e^{\frac{\gamma}{1-\alpha} k^{1-\alpha}} \end{aligned}$$

The function $x \longmapsto x^{-2\beta} e^{\frac{\gamma}{1-\alpha} x^{1-\alpha}}$ being increasing for $x \geq c_{\gamma,\alpha}$, we then obtain, considering an integer $t > c_{\gamma,\alpha}$:

$$\sum_{k=1}^{n} \gamma_k^2 \prod_{l=k+1}^{n} (1-\gamma_l)^2 \leq \gamma^2 e^{-2\Gamma_n} \left(C_t + \int_t^n x^{-2\alpha} e^{\frac{\gamma}{1-\beta} x^{1-\alpha}} dx \right).$$

We can write $x^{-2\alpha} e^{Kx^{1-\alpha}} = \left(e^{Kx^{1-\alpha}} \right)' x^{-\alpha} K^{-1} (1-\alpha)^{-1}$ and integrating by parts, we obtain for a large enough n :

$$\sum_{k=1}^{n} \gamma_k^2 \prod_{l=k+1}^{n} (1-\gamma_l)^2 \leq \gamma^2 e^{-2\Gamma_n} \left(C_t + \frac{e^{2\Gamma_n}}{2\gamma} n^{-\alpha} \right) \leq 2\gamma_n.$$

$\diamond$

We now study the situation where $\alpha = 1$ and write (using the monotonicity of the logarithm function) :

$$\gamma \log(n+1)/(k+1) \leq \sum_{k+1}^{n} \gamma_l \leq \gamma \log(n/k)$$

In that case, we obtain

$$\begin{aligned} \sum_{k=1}^{n} \gamma_k^2 \prod_{l=k+1}^{n} (1-\gamma_l)^2 &\leq \gamma^2 \sum_{k=1}^{n} k^{-2} e^{-2\gamma \log((n+1)/(k+1))} \\ &\leq (n+1)^{-2\gamma} \gamma^2 \sum_{k=1}^{n} k^{-2} (k+1)^{2\gamma}. \end{aligned}$$

Two situations are now possible : if $2\gamma - 2 \neq -1$, then we have

$$\sum_{k=1}^{n} k^{-2} (k+1)^{2\gamma} \leq n^{-2\gamma+1},$$

so that

$$\mathbb{V}_n \leq \frac{2\sigma^2 \gamma^2}{n},$$

for n large enough. Now, if oppositely $2\gamma - 2 = -1$, then $\gamma = 1/2$ and in that case for n large enough :

$$\mathbb{V}_n \leq \sigma^2 \frac{\log n}{2n}$$

Global rate We shall end this introductory section by an upper bound on the global rate of convergence of the SGD in the linear case, when $\gamma_n = \gamma n^{-\alpha}$. We have

— When $\alpha < 1$, then

$$\mathbb{B}_n + \mathbb{V}_n \lesssim |\theta_0 - m|^2 e^{-2\gamma(1-\alpha)^{-1} n^{1-\alpha}} + 2\sigma^2 \gamma_n,$$

meaning that the bias term is exponentially fast decreasing while the variance term is much more slower of size $\sigma^2 n^{-\alpha}$.

— When $\alpha = 1$, we can see that the bias term is much slower than in the previous situation, and decreases as $n^{-2\gamma}$ while the variance term is $\sigma^2 \log n/n$ when $\gamma = 1/2$ and $\sigma^2 n^{-1}$ otherwise.

Hence, the global rate is in each situation slower than n^{-1} (which is of course the optimal rate in this toy example). Moreover :

— we attain this rate only when $\alpha = 1$ and γ is sufficiently large (strictly larger than $1/2$).

— When $\alpha = 1$ and $\gamma = 1/2$, the rate is slightly deteriorated into $\log n/n$.

— When $\alpha < 1$, the rate is seriously damaged into $n^{-\alpha}$.

To sum up, we have shown :

Theorem 3.1.1 *Assume $\gamma_n = \gamma n^{-\alpha}$ with $\alpha < 1$, then the quadratic risk is dominated by the variance term and*

$$\mathbb{B}_n + \mathbb{V}_n \lesssim \sigma^2 n^{-\alpha}.$$

Assume $\gamma_n = \gamma n^{-1}$, then the quadratic risk is upper bounded by

$$\mathbb{B}_n + \mathbb{V}_n \lesssim \log(n)^{\mathbf{1}_{\gamma=1/2}} n^{-(1\wedge 2\gamma)}$$

3.1.3 General linear one-dimensional function

We consider now the slightly different situation where $f(x) = \mu x$. All the upper bounds have to be slightly modified. Nevertheless, the modifications are very mild and it is straightforward to check that

— If $\alpha < 1$, then

$$\mathbb{B}_n \lesssim |\theta_0 - m|^2 e^{-2\gamma\mu(1-\alpha)^{-1}n^{1-\alpha}}$$

and

$$\mathbb{V}_n \lesssim 2\sigma^2\gamma_n.$$

— If $\alpha = 1$, then the bias term is now upper bounded by

$$\mathbb{B}_n \lesssim |\theta_0 - m|^2 n^{-2\gamma\mu},$$

while the variance term is treated as follows. If $\gamma\mu \neq 1/2$ then

$$\mathbb{V}_n \lesssim \sigma^2 n^{-1},$$

while the rate is deteriorated with a log term when $\gamma\mu = 1/2$.

Again, the best performances are attained with $\gamma_n = \gamma n^{-1}$ but these performances highly depend on the size of $\gamma\mu - 1/2$. Later, we will discuss on this non adaptive issue of the SGD.

To sum-up, we have the next convergence rate result.

Theorem 3.1.2 *Assume $\gamma_n = \gamma n^{-\alpha}$ with $\alpha < 1$, then the quadratic risk is dominated by the variance term and*

$$\mathbb{B}_n + \mathbb{V}_n \lesssim \sigma^2 n^{-\alpha}.$$

Assume $\gamma_n = \gamma n^{-1}$, then the quadratic risk is upper bounded by

$$\mathbb{B}_n + \mathbb{V}_n \lesssim \log(n)^{\mathbf{1}_{\mu\gamma=1/2}} n^{-(1\wedge 2\mu\gamma)}$$

We should end this paragraph by saying that it is also possible to handle general linear functions up to a change of basis. We will not discuss in details this case and now switch to a more general settings.

3.2 Rate of SGD for general convex function

We begin with the general situation of stochastic gradient descent when we handle a convex minimization problem. Hence, f is assumed to be convex, but not necessarily strongly convex. Moreover, we assume that the minimization problem is localized on a (possibly large) euclidean ball of $\mathbb{R}^p$ of radius D :

$$\mathcal{C} := \{\theta \in \mathbb{R}^p \,:\, \|\theta\|_2 \leq D\}\,.$$

Such assumption is not necessarily strong because from a practical point of view, we have in general an idea of a reasonnable D such that the minimizer of f over $\mathbb{R}^p$ is located inside $\mathcal{C}$. We then handle a projected stochastic gradient descent defined as

$$\theta_n := \Pi_{\mathcal{C}}\left(\theta_{n-1} - \gamma_n \nabla f(\theta_{n-1}) + \gamma_n \Delta M_n\right),$$

where $\Pi_{\mathcal{C}}$ refers to the euclidean projection on the sphere of radius D. We will establish the next theorem.

Theorem 3.2.1 *Assume that $(\theta_n)_{n\geq 1}$ is a projected SGD in $B(0,D)$ and that*

$$\|\Delta M_n + \nabla f(\theta_{n-1})\|_\infty \leq B.$$

Then the choice $\gamma_n = \frac{2D}{B\sqrt{n}}$ yields

$$\mathbb{E} f\left(\bar{\theta}_n\right) - f(\theta^\star) \leq \frac{2DB}{\sqrt{n}},$$

where

$$\bar{\theta}_n := \frac{1}{n}\sum_{k=0}^{n-1} \theta_k.$$

<u>*Proof :*</u> We show the proof of this result by repeting almost exactly the proof of the convergence rate when we handle the deterministic case. To alleviate the proof, we replace the noisy gradient evaluation by the following notation :

$$\nabla f(\theta_{n-1}) - \Delta M_n = f_n'(\theta_{n-1}),$$

where f_n' is a noisy gradient centered around the true value $\nabla f(\theta_{n-1})$.

$$\begin{aligned}
\|\theta_n - \theta^\star\|^2 &= \|\Pi_{\mathcal{C}}\left(\theta_{n-1} - \gamma_n f_n'(\theta_{n-1}) - \theta^\star\right)\|^2 \\
&\leq \|\theta_{n-1} - \gamma_n f_n'(\theta_{n-1}) - \theta^\star\|^2 \\
&\leq \|\theta_{n-1} - \theta^\star\|^2 + B^2\gamma_n^2 - 2\gamma_n \langle \theta_{n-1} - \theta^\star, f_n'(\theta_{n-1})\rangle.
\end{aligned}$$

Now, we can take the expectation with respect to $\mathcal{F}_{n-1}$ and deduce that

$$\mathbb{E}\left[\|\theta_n - \theta^\star\|^2 | \mathcal{F}_{n-1}\right] \leq \|\theta_{n-1} - \theta^\star\|^2 + B^2\gamma_n^2 - 2\gamma_n \langle \theta_{n-1} - \theta^\star, \nabla f(\theta_{n-1})\rangle.$$

Now, the convex property $f(x) + \langle \nabla f(x), y - x\rangle \leq f(y)$ yields

$$\mathbb{E}\left[\|\theta_n - \theta^\star\|^2 | \mathcal{F}_{n-1}\right] \leq \|\theta_{n-1} - \theta^\star\|^2 + B^2\gamma_n^2 - 2\gamma_n\left[f(\theta_{n-1}) - f(\theta^\star)\right].$$

If we now compute the whole expectation, we then obtain

$$\mathbb{E}\|\theta_n - \theta^\star\|^2 \leq \mathbb{E}\|\theta_{n-1} - \theta^\star\|^2 + B^2\gamma_n^2 - 2\gamma_n\left[\mathbb{E} f(\theta_{n-1}) - f(\theta^\star)\right].$$

An equivalent writing is

$$\mathbb{E}f(\theta_{n-1}) - f(\theta^\star) \leq \frac{B^2\gamma_n}{2} + \frac{1}{2\gamma_n}\left[\mathbb{E}\|\theta_{n-1} - \theta^\star\|^2 - \mathbb{E}\|\theta_n - \theta^\star\|^2\right].$$

We now use a telescopic summation argument and obtain

$$\sum_{k=1}^{n} \mathbb{E}f(\theta_{k-1}) - f(\theta^\star) \leq \sum_{k=1}^{n} \frac{B^2\gamma_k}{2} + \sum_{k=1}^{n} \frac{1}{2\gamma_k}\left[\mathbb{E}\|\theta_{k-1} - \theta^\star\|^2 - \mathbb{E}\|\theta_k - \theta^\star\|^2\right].$$

The rough bound $\mathbb{E}\|\theta_{k-1} - \theta^\star\|^2 \leq 4D^2$ then leads to

$$\sum_{k=1}^{n} \mathbb{E}f(\theta_{k-1}) - f(\theta^\star) \leq \sum_{k=1}^{n} \frac{B^2\gamma_k}{2} + \frac{4D^2}{2\gamma_n} n \leq 2DB\sqrt{n} \qquad \text{when} \qquad \gamma_n = 2DB\sqrt{n}.$$

Now, using convexity, we deduce that

$$\mathbb{E}f\left(\bar{\theta}_{n-1}\right) - f(\theta^\star) \leq \frac{1}{n}\sum_{k=1}^{n}\left(\mathbb{E}f(\theta_{k-1}) - f(\theta^\star)\right) \leq \frac{2DB}{\sqrt{n}}.$$

This ends the proof of the result. □

We should end the paragraph with the following remark. The bound attained by the SGD algorithm in the convex situation (without any assumption on strong convexity) is minimax optimal. It can be shown that any first order method with noisy gradients cannot attain the minimum of a convex function faster than $n^{-1/2}$. This result may be found in the seminal contribution of Nemirovski and Yudin (1983).

At last, we should remark that the gain of the algorithm (the step-size sequence) needs a very specific calibration, with $\gamma_n \propto n^{-1/2}$. Moreover, γ_n has to be linked with B and D. If the dependence with D is not annoying, it is not exactly the same problem with its dependence with respect to B, which is in general unknown. This last dependency could be avoided, with additional technicalicities.

3.3 Rate of SGD for strongly convex function

The study of the strongly convex situation is motivated by at least two important facts.

— First, we have seen in Chapter 1 that some great improvements could be obtained with a strongly convex function (instead of the simplest convex property), leading to a linear rate of convergence instead of a polynomial one, even with the gradient descent algorithm.

— Second, in the introduction, we have seen that the SGD algorithm with a linear drift ($\nabla f(x) = \mu x$) may attain some better convergence rate than only $n^{-1/2}$. The key feature of this linear case is indeed to belong to the class of strongly convex minimization problem.

We then study the SGD with a μ strongly convex function f and show the next result.

Theorem 3.3.1 *Let $(\theta_n)_{n\geq 1}$ defined by*

$$\theta_n = \Pi_D\left(\theta_{n-1} - \gamma_n f_n'(\theta_{n-1})\right),$$

then :

i) If $\gamma_n = \frac{2}{\mu(n+1)}$, we have

$$\mathbb{E}f\left(\frac{2}{n(n+1)}\sum_{k=1}^{n} k\theta_{k-1}\right) - f(\theta^\star) \leq \frac{2B^2}{\mu(n+1)}$$

ii) If $\gamma_n = \frac{2}{\mu(n+1)}$, *we have*

$$\mathbb{E}f\left(\frac{1}{n}\sum_{k=1}^{n}\theta_{k-1}\right) - f(\theta^\star) \leq \frac{B^2\log(n)}{2n\mu}.$$

<u>*Proof :*</u> <u>Proof of *i)* :</u> We first deal with the step-size $\gamma_n = \frac{2}{\mu(n+1)}$.

We begin the proof as usual, comparing $\|\theta_n - \theta^\star\|^2$ with $\|\theta_{n-1} - \theta^\star\|^2$:

$$\begin{aligned}\|\theta_n - \theta^\star\|^2 &= \|\Pi_{\mathcal{C}}\left(\theta_{n-1} - \gamma_n f_n'(\theta_{n-1}) - \theta^\star\right)\|^2 \\ &\leq \|\theta_{n-1} - \gamma_n f_n'(\theta_{n-1}) - \theta^\star\|^2\end{aligned}$$

The two first inequalities are standard (1-Lipschitz property). Computing now the expectation with respect to $\mathcal{F}_{n-1}$ yields

$$\mathbb{E}\left[\|\theta_n - \theta^\star\|^2 |\mathcal{F}_{n-1}\right] \leq \|\theta_{n-1} - \theta^\star\|^2 + B^2\gamma_n^2 - 2\gamma_n\left[f(\theta_{n-1}) - f(\theta^\star) + \frac{\mu}{2}\|\theta_{n-1} - \theta^\star\|^2\right].$$

The last line is derived from the strong convexity property :

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

which is applied with $x = \theta_{n-1}$ and $y = \theta^\star$. If we now follow the same strategy and move the difference $f(\theta_{n-1}) - f(\theta^\star)$ on the left hand side, we deduce that

$$\mathbb{E}\left[f(\theta_{n-1} - f(\theta^\star)|\mathcal{F}_{n-1}\right] \leq \frac{B^2\gamma_n}{2} + \frac{1}{2}\left(\gamma_n^{-1} - \mu\right)\|\theta_{n-1} - \theta^\star\|^2 - \frac{\gamma_n^{-1}}{2}\|\theta_n - \theta^\star\|^2$$

Now, we shall use the definition of γ_n and check that

$$\frac{1}{2}\left(\gamma_n^{-1} - \mu\right) = \frac{\mu(n-1)}{4} \qquad \text{and} \qquad \frac{\gamma_n^{-1}}{2} = \frac{\mu(n+1)}{4}.$$

Then, we get that

$$\mathbb{E}\left[f(\theta_{n-1} - f(\theta^\star)|\mathcal{F}_{n-1}\right] \leq \frac{B^2\gamma_n}{2} + \frac{\mu(n-1)}{4}\|\theta_{n-1} - \theta^\star\|^2 - \frac{\mu(n+1)}{4}\|\theta_n - \theta^\star\|^2.$$

Summing the above terms with a weights proportional to n yields

$$\begin{aligned}\sum_{k=1}^{n} k\mathbb{E}f(\theta_{k-1} - f(\theta^\star) &\leq \sum_{k=1}^{n}\frac{B^2k}{\mu(k+1)} + \frac{1}{4}\sum_{k=1}^{n}k(k-1)\|\theta_{k-1} - \theta^\star\|^2 - \frac{1}{4}\sum_{k=1}^{n}k(k+1)\|\theta_k - \theta^\star\|^2 \\ &\leq \frac{B^2}{\mu}n + \frac{1}{4}\left[0 - n(n+1)\mathbb{E}\|\theta_n - \theta^\star\|^2\right] \\ &\leq \frac{B^2}{\mu}n\end{aligned}$$

Again, using the convexity of f, we are led to

$$\mathbb{E}f\left(\frac{2}{n(n+1)}\sum_{k=1}^{n}k\theta_{k-1}\right) - f(\theta^\star) \leq \frac{2B^2}{\mu(n+1)}$$

$\diamond$

Proof of $ii)$: We now deal with $\gamma_n = \frac{1}{n\mu}$ and repeat the arguments above : we are led to

$$\mathbb{E}\left[f(\theta_{n-1} - f(\theta^\star)\,|\mathcal{F}_{n-1}\right] \leq \frac{B^2\gamma_n}{2} + \frac{\mu(n-1)}{2}\|\theta_{n-1} - \theta^\star\|^2 - \frac{\mu n}{2}\|\theta_n - \theta^\star\|^2.$$

Summing these inequalities from 1 to n, we obtain

$$\begin{aligned}\sum_{k=1}^n k\mathbb{E}f(\theta_{k-1} - f(\theta^\star) &\leq \sum_{k=1}^n \frac{B^2}{2\mu k} + \frac{1}{2}\sum_{k=1}^n (k-1)\|\theta_{k-1} - \theta^\star\|^2 - \frac{1}{2}\sum_{k=1}^n k\|\theta_k - \theta^\star\|^2 \\ &\leq \frac{B^2\log(n)}{2n\mu}.\end{aligned}$$

□

Even though the bound is better in $i)$ for large values of n (because of the log term), it should however be pointed out that the second point is better for smaller values of n.

Note that it is also possible to deduce a result for SGD on μ strongly convex function when the noise associated to the martingale increment has a bounded variance and when f is L-smooth. The key descent lemma becomes

$$\mathbb{E}\left[\|\theta_n - \theta^\star\|^2\,|\mathcal{F}_{n-1}\right] \leq \|\theta_{n-1} - \theta^\star\|^2 + \sigma^2\gamma_n^2 + \frac{L}{2}\gamma_n^2\|\theta_{n-1} - \theta^\star\|^2 - 2\mu\gamma_n\|\theta_{n-1} - \theta^\star\|^2.$$

If we define now $U_n = \|\theta_n - \theta^\star\|^2$, the inequality above leads to

$$\mathbb{E}U_n \leq \mathbb{E}U_{n-1}\left(1 - 2\gamma_n\mu + \frac{L}{2}\gamma_n^2\right) + \gamma_n^2\sigma^2.$$

This last inequality may be used recursively to obtain a similar upper bound on the rate of convergence of U_n. In particular, the results derived in Section 3.1 still hold with slightly different multiplicative constants.

3.4 Deviation inequalities

This last paragraph is devoted to the analysis of other theoretical results on stochastic gradient algorithm from a non-asymptotic point of view. In particular, it is possible to go beyond the quadratic risk : a first kind of result could be an analysis of the $\mathbb{L}^p$ loss. A second kind of result consists in providing deviation inequalities (high probability bounds) on the stochastic algorithm, and deriving upper bounds of the form

$$\mathbb{P}\left(\|\theta_n - \theta^\star\| \geq \epsilon\right) \leq e^{-r(n,\epsilon)}$$

where $r(n,\epsilon)$ is a suitable quantity that should be related to the central limit theorem we will obtain in the next chapter.

Assume that $(\theta_n)_{n\geq 1}$ is a strongly convex SGD while f is μ-strongly convex used with a step size $(\gamma_n)_{n\geq 1}$. Below, we denote $\epsilon = 1$ if $\gamma_n = \gamma_1 n^{-1}$ when $\gamma_1\mu = 1/2$. Otherwise, $\epsilon = 0$.

Theorem 3.4.1 *Assume that $(\theta_n)_{n\geq 1}$ is a strongly convex SGD while f is μ-strongly convex. Then we have :*

— *If $\alpha < 1$:*

$$\mathbb{P}\left(\|\theta_n - \theta^\star\|^2 \geq \|\theta_0 - \theta^\star\|^2 e^{-2\mu\Gamma_n} + \delta\right) \leq e^{-\frac{\delta^2}{\sigma^2\gamma_n}}.$$

— *If* $\alpha = 1$ *:*

$$\mathbb{P}\left(\|\theta_n - \theta^\star\|^2 \geq \|\theta_0 - \theta^\star\|^2 e^{-2\mu\Gamma_n} + \delta\right) \leq e^{-\frac{\delta^2}{\sigma^2 n^{-(\gamma_1\mu\wedge 1)}\log(n)^\epsilon}}.$$

Without loss of generality, we assume below that $\theta^\star = 0$.

It is well known that some concentration bounds may be obtained with the help of the Laplace transform, following the so-called Chernoff method. We follow a simplified framework of SGD evolution as introduced in Woodroofe (1972). We define our stochastic algorithm as

$$\theta_{n+1} = \theta_n - \gamma_{n+1} f_n'(\theta_n),$$

and introduce the Laplace transform of the algorithm

$$\phi_n(t) := \mathbb{E}[e^{t\|\theta_n - \theta^\star\|^2}].$$

To obtain some good deviation inequalities on $\theta_n - \theta^\star$, we need to find an upper bound of ϕ_{X_n} for a suitable value of t. For this purpose, we still use the successive conditionning argument and remark that

$$\phi_n(t) = \mathbb{E}[e^{t\|\theta_n - \theta^\star\|^2}] = \mathbb{E}\left[\mathbb{E}\left[e^{t\|\theta_n - \theta^\star\|^2}\right]|\mathcal{F}_{n-1}\right]$$

Now, we shall write

$$e^{t\|\theta_n - \theta^\star\|^2} = e^{t\|\theta_{n-1} - \theta^\star - \gamma_n f_n'(\theta_{n-1})\|^2} = e^{t\|\theta_{n-1} - \theta^\star - \gamma_n \nabla f(\theta_{n-1}) + \gamma_n \Delta M_n\|^2}$$

We use the following computation :

$$\begin{aligned}
\|\theta_n - \theta^\star\|^2 &= \|\theta_{n-1} - \theta^\star\|^2 + \gamma_n^2\|\Delta M_n - \nabla f(\theta_{n-1})\|^2 - 2\gamma_n\langle \nabla f(\theta_{n-1}), \theta_{n-1} - \theta^\star\rangle \\
&\quad + 2\gamma_n\langle \theta_{n-1} - \theta^\star, \Delta M_n\rangle \\
&\leq \|\theta_{n-1} - \theta^\star\|^2 + \gamma_n^2\|\Delta M_n - \nabla f(\theta_{n-1})\|^2 - \mu\gamma_n\|\theta_{n-1} - \theta^\star\|^2 \\
&\quad + 2\gamma_n\langle \theta_{n-1} - \theta^\star, \Delta M_n\rangle
\end{aligned}$$

where we used the strong convexity property in the last line. Now, we turn back to the Laplace transform computation, conditionned to step $n-1$:

$$\begin{aligned}
\mathbb{E}[e^{t\|\theta_n - \theta^\star\|^2}] &= \mathbb{E}\left[\mathbb{E}\left[e^{t\|\theta_n - \theta^\star\|^2}\right]|\mathcal{F}_{n-1}\right] \\
&\leq \mathbb{E}\left[e^{t(1-\mu\gamma_n)\|\theta_{n-1} - \theta^\star\|^2 + t\gamma_n^2 B^2}\right. \\
&\quad \left.\times\, \mathbb{E}\left[e^{2t\gamma_n\langle \theta_{n-1} - \theta^\star, \Delta M_n\rangle}\,|\,\mathcal{F}_{n-1}\right]\right]
\end{aligned}$$

where we used the boundedness of the sequence $(\theta_n)_{n\geq 1}$ in $B(0, D)$ and of the increment : $\|\Delta M_n + \nabla f(\theta_{n-1})\| \leq B$. The key observation now comes from an application of the Hoeffding Lemma (see Lemma 3.4.1). In particular, the Cauchy Schwarz inequality leads to

$$|\langle \theta_{n-1} - \theta^\star, \Delta M_n\rangle| \leq 2DB.$$

Therefore, we obtain that

$$\mathbb{E}\left[e^{2t\gamma_n\langle \theta_{n-1} - \theta^\star, \Delta M_n\rangle}\,|\,\mathcal{F}_{n-1}\right] \leq e^{\frac{4t^2\gamma_n^2\{2DB\}^2}{8}} = e^{t^2\gamma_n^2 D^2 B^2}.$$

Consequently, we obtain the key recursive upper-bound :

$$\phi_n(t) \leq \phi_{n-1}(t(1-\gamma_n\mu))e^{t^2\gamma_n^2 D^2(B^2+2)}.$$

This inequality leads to

$$\phi_n(t) \leq \phi_0 \left(t \prod_{k=1}^{n} (1 - \gamma_k \mu) \right) e^{t^2 D^2 (B^2+2) \sum_{k=1}^{n} \gamma_k^2}.$$

If we now introduce the notation

$$\pi_k := \prod_{i=1}^{k} (1 - \gamma_i \mu),$$

we can deduce that

$$\phi_n(t) \leq e^{t\|\theta_0 - \theta^\star\|^2 \pi_n + t^2 D^2 (B^2+2) S \sum_{k=1}^{n} \pi_n^2 \pi_k^{-2} \gamma_k^2}.$$

The last sum can be studied with a series/integral comparison argument (see the beginning of this chapter). If $\alpha < 1$, then we have shown that

$$\sum_{k=1}^{n} \pi_n^2 \pi_k^{-2} \gamma_k^2 \lesssim \pi_n^2 \gamma_1^2 \int_1^n x^{-2\alpha} e^{2\mu\gamma_1 x^{1-\alpha}} \lesssim \gamma_n$$

while if $\alpha = 1$ we have

$$\sum_{k=1}^{n} \pi_n^2 \pi_k^{-2} \gamma_k^2 \lesssim \pi_n^2 \gamma_1^2 \int_1^n x^{-2} e^{2\mu\gamma_1 \log(x)} \lesssim \gamma_n \log(n)^\epsilon,$$

where $\epsilon = \mathbf{1}_{\gamma\mu = 1/2}$. In any case, we have shown the next upper bound of the Laplace transform (that can be made more explicit up to additional painful technicalities). If $\alpha < 1$:

$$\phi_n(t) \leq e^{t\|\theta_0 - \theta^\star\|^2 \pi_n + \sigma^2 t^2 \gamma_n}. \tag{3.2}$$

while if $\alpha = 1$:

$$\phi_n(t) \leq e^{t\|\theta_0 - \theta^\star\|^2 \pi_n + \sigma^2 t^2 n^{-(1 \wedge \gamma_1 \mu)} \log(n)^\epsilon}. \tag{3.3}$$

In particular, we can immediately see that the term $t\|\theta_0 - \theta^\star\|^2 \pi_n$ represents the Bias term involved in the evolution of the SGD while the t^2 term is the variance one. We only deal with the case $\alpha < 1$. The Chernoff method applied with Equation Equation (3.2) yields

$$\begin{aligned}
\mathbb{P}\left(\|\theta_n - \theta^\star\|^2 > \|\theta_0 - \theta^\star\|^2 \pi_n + \delta\right) &\leq e^{t\|\theta_n - \theta^\star\|^2} e^{-t\|\theta_0 - \theta^\star\|^2 \pi_n - t\delta} \\
&\leq e^{t\|\theta_0 - \theta^\star\|^2 \pi_n + \sigma^2 t^2 \gamma_n} e^{-t\|\theta_0 - \theta^\star\|^2 \pi_n - t\delta} \\
&\leq e^{\sigma^2 t^2 \gamma_n - t\delta}
\end{aligned}$$

Optimizing t in the above upper bound leads to the choice :

$$t_n = \frac{\delta}{2\sigma^2 \gamma_n}.$$

Up to this last choice, we deduce that

$$\mathbb{P}\left(\|\theta_n - \theta^\star\|^2 > \|\theta_0 - \theta^\star\|^2 \pi_n + \delta\right) \leq e^{-\frac{\delta^2}{2\sigma^2 \gamma_n}}.$$

This ends the proof. □

Lemma 3.4.1 *Assume that Y is a bounded and centered real random variable such that almost surely $c \leq Y \leq d$. Then,*

$$\forall s \geq 0 \qquad E[e^{sY}] \leq e^{\frac{s^2 (d-c)^2}{8}}.$$

We should end the Chapter with a last observation. Clearly, the deviation result above leads to a sub-Gaussian behaviour of the algorithm around $\theta^\star$ at a scale $\sqrt{\gamma}_n$. Hence, it would be natural to obtain a central limit theorem (with this rate $\sqrt{\gamma}_n$) for the *rescaled algorithm* :

$$\hat{\theta}_n = \gamma_n^{-1/2} \left(\theta_n - \theta^\star\right).$$

This is the purpose of the next chapter.

Chapitre 4

Central limit theorem

4.1 Motivation

In this chapter, we study the stochastic gradient descent algorithm defined by a step-size sequence $(\gamma_n)_{n\geq 1}$ with a μ-strongly convex function f. We have shown that when $\gamma_n = \gamma_1 n^{-\alpha}$, the SGD satisfies

— When $\alpha < 1$

$$\mathbb{E}\|\theta_{n+1} - \theta^\star\|^2 \lesssim \frac{1}{n^\alpha},$$

— When $\alpha = 1$ and $\gamma_1\mu > 1/2$

$$\mathbb{E}\|\theta_{n+1} - \theta^\star\|^2 \lesssim \frac{1}{n},$$

— When $\alpha = 1$ and $\gamma_1\mu = 1/2$

$$\mathbb{E}\|\theta_{n+1} - \theta^\star\|^2 \lesssim \frac{\log n}{n},$$

— When $\alpha = 1$ and $\gamma_1\mu < 1/2$

$$\mathbb{E}\|\theta_{n+1} - \theta^\star\|^2 \lesssim \frac{1}{n^{2\gamma_1\mu}},$$

4.2 Rescaling a stochastic algorithm

In this paragraph, we establish a (functional) Central Limit Theorem when the S.G.D. algorithm defined by

$$X_n = X_{n-1} - \gamma_n f_n'(X_{n-1}). \tag{4.1}$$

when f is a μ strongly convex function and when $(\mathbf{H}_{SC}(\mu, L))$ holds.

• Assumption $(\mathbf{H}_{SC}(\mu, L))$: f is μ strongly convex with L-Lipschitz gradient.

In particular, f admits a unique minimum $x^\star$. Without loss of generality, we assume that $x^\star = 0$.

We also introduce an additional assumption on the noise sequence of the martingale increment. We assume that this noise has a bounded moment that may be related to the size of f :

• Assumption $(\mathbf{H}_{\boldsymbol{\sigma},\boldsymbol{p}})$: $(p \geq 1)$ For any integer n, we have :

$$\mathbb{E}(\|\Delta M_{n+1}\|^p|\mathcal{F}_n) \leq \sigma^2(1 + f(X_n))^p.$$

4.2.1 Definition of the rescaled process

We start with an appropriate rescaling by a factor $\sqrt{\gamma_n}$. More precisely, we define a sequence $(\check{X}_n)_{n\geq 1}$:

$$\check{X}_n = \frac{X_n(-x^\star)}{\sqrt{\gamma_n}}$$

Given that f is $\mathcal{C}^2$ (and that $x^\star = 0$), we "linearize" ∇f around 0 with a Taylor formula and obtain that $\xi_n \in [0, X_n]$ exists such that :

$$\nabla f(X_n) = D^2 f(\xi_n) X_n.$$

Therefore, we can compute that :

$$\check{X}_{n+1} = \check{X}_n + \gamma_{n+1} b_n(\check{X}_n) + \sqrt{\gamma_{n+1}} \Delta M_{n+1}$$

where b_n is defined by :

$$\forall z \in \mathbb{R}^p \qquad b_n(z) = \frac{1}{\gamma_{n+1}} \left(\sqrt{\frac{\gamma_n}{\gamma_{n+1}}} - 1 \right) z + \bar{C}_n z, \quad z \in \mathbb{R}^{2d}, \tag{4.2}$$

where :

$$\bar{C}_n := -\sqrt{\frac{\gamma_n}{\gamma_{n+1}}} D^2 f(\xi_n). \tag{4.3}$$

It is important to observe that if $\gamma_n = \gamma_1 n^{-\alpha}$, then

$$\frac{1}{\gamma_{n+1}} \left(\sqrt{\frac{\gamma_n}{\gamma_{n+1}}} - 1 \right) = \gamma^{-1}(n+1)^\alpha \left[1 + \frac{\alpha}{2n} + o(n^{-1}) - 1 \right] = \begin{cases} o(n^{\beta-1}) & \text{if} \quad \alpha < 1 \\ \frac{1}{2\gamma} + o(1) & \text{if} \quad \alpha = 1 \end{cases} \tag{4.4}$$

4.2.2 Interpolated continuous-time process

We associate to the sequence $(\check{X}_n)_{n\geq 1}$ a sequence $(\bar{X}^{(n)})_{n\geq 1}$ of continuous-time processes. In this view, it is convenient to introduce the following standard notation.

Definition 4.2.1 (Time interpolation) *For any integer n and any time $t > 0$, we define*

$$N(n,t) := \min \left\{ m \geq n, \sum_{k=n+1}^{m} \gamma_k > t \right\}$$

and

$$\underline{t}_n := \Gamma_{N(n,t)} - \Gamma_n$$

where

According to this time interpolation (switched with Γ_n), we define the continuous-time process.

Definition 4.2.2 (Process $\bar{X}_t^{(n)}$)

$$\bar{X}_t^{(n)} = \check{X}_n + B_t^{(n)} + M_t^{(n)}, \quad t \geq 0, \tag{4.5}$$

where :

$$B_t^{(n)} = \sum_{k=n+1}^{N(n,t)} \gamma_k b_{k-1}(\check{X}_{k-1}) + (t - \underline{t}_n) b_{N(n,t)}(\check{X}_{N(n,t)}),$$

and

$$M_t^{(n)} = \sum_{k=n+1}^{N(n,t)} \sqrt{\gamma_k} \Delta M_k + \sqrt{t - \underline{t}_n} \Delta M_{N(n,t)+1}.$$

To obtain a CLT, we show that $(\bar{X}^{(n)})_{n\geq 1}$ converges in distribution to a stationary diffusion, following a classical roadmap based on a tightness result and on an identification of the limit as a solution to a martingale problem.

4.3 Tightness of $(\bar{X}^{(n)})_{n\geq 1}$

Tightness of a sequence of stochastic process is a useful tool to obtain convergence results, in particular convergence in distribution. It is well known that tightness is a compactness criterion on certain path-space, which is a preliminary step before using a complementary identifiability step, for proving weak convergence.

Here, the sequence of processes $(\bar{X}^{(n)})_{n\geq 0}$ is built with continuous time process. A classical criterion (see, *e.g.*, Theorem 8.3 of Billingsley) shows that a sufficient condition for the tightness of $(\bar{X}^{(n)})_{n\geq 1}$ (for the weak topology induced by the uniform convergence on compacts intervals) is the following property.

Proposition 4.3.1 (Tightness criterion on $\mathcal{C}$) *Assume that for any $T > 0$, for any positive ε and η, a $\delta > 0$ exists and an integer n_0 exists such that :*

$$\forall t \in [0,T] \quad \forall n \geq n_0 \qquad \mathbb{P}(\sup_{s\in[t,t+\delta]} \|\bar{X}_s^{(n)} - \bar{X}_t^{(n)}\| \geq \varepsilon) \leq \eta\delta.$$

Then, the sequence $(\bar{X}^{(n)})_{n\geq 1}$ is tight for the weak topology induced by the weak convergence on compact intervals.

The next lemma holds for any sequence of processes that satisfy Equation (4.5).

Lemma 4.3.1 *Assume that :*
- *f satisfies $(\mathbf{H}_{SC}(\mu, L))$,*
- *$\sup_{k\geq 1} \mathbb{E}[\|\check{X}_k\|^2] < +\infty$,*
- *a $p > 2$ exists such that $\sup_{k\geq 1} \mathbb{E}[\|\Delta M_k\|^p] < +\infty$.*

Then $(\bar{X}^{(n)})_{n\geq 1}$ is tight (for the weak topology induced by the weak convergence on compact intervals).

Proof : First, note that $\bar{X}_0^{(n)} = \check{X}_n$, the assumption $\sup_{k\geq 1} \mathbb{E}[\|\check{X}_k\|^2] < +\infty$ implies the tightness of $(\bar{X}_0^{(n)})_{n\geq 1}$ because we have :

$$\lim_{K\longrightarrow+\infty} \limsup_n \mathbb{P}\left(\|\bar{X}_0^{(n)}\| \geq K\right) = 0.$$

We consider $B^{(n)}$ and $M^{(n)}$ separately and begin by the drift term $B^{(n)}$. On the one hand,

$$\mathbb{P}\left(\sup_{s\in[t,t+\delta]} \|B_s^{(n)} - B_t^{(n)}\| \geq \varepsilon\right) \leq \mathbb{P}\left(\sum_{k=N(n,t)}^{N(n,t+\delta)+1} \gamma_k \|b_{k-1}(\check{X}_{k-1})\| \geq \varepsilon\right).$$

We known that $\|b_k(z)\| \leq C(1 + \|z\|)$ for a universal constant C independent on k because f is L-smooth, which implies that $\|D^2 f\| \leq L$. Now, consider non negative ϵ and η, the Chebyschev inequality yields :

$$\mathbb{P}\left(\sup_{s\in[t,t+\delta]} \|B_s^{(n)} - B_t^{(n)}\| \geq \varepsilon\right) \leq \varepsilon^{-2}\mathbb{E}\left[\left(\sum_{k=N(n,t)}^{N(n,t+\delta)+1} \gamma_k(1 + \|(\check{X}_{k-1})\|)\right)^2\right]$$

The Jensen inequality (on the supremum) and the fact that $\sum_{k=N(n,t)}^{N(n,t+\delta)+1} \gamma_k \leq 2\delta$ when n is large enough imply that a constant C exists such that for large enough n and for a small enough δ :

$$\mathbb{P}\left(\sup_{s\in[t,t+\delta]} \|B_s^{(n)} - B_t^{(n)}\| \geq \varepsilon\right) \leq \varepsilon^{-2} \times C\delta^2(1 + \sup_{k\geq 1} \mathbb{E}[\|\check{X}_k\|^2]) \leq \eta\delta,$$

if we choose $\delta = \epsilon^2 C^{-1}\eta^{-1}$. ◇

We now consider the martingale component $M^{(n)}$: if we denote $\tau = \sqrt{\frac{t-t_n}{\gamma_{N(n,t)+1}}}$, we have for any $s \geq 0$,

$$M_s^{(n)} = (1-\tau)M_{N(n,s)}^{(n)} + \tau M_{N(n,s)+1}^{(n)}$$

so that $\|M_s^{(n)} - M_t^{(n)}\| \leq \max\{\|M_{N(n,s)}^{(n)} - M_t^{(n)}\|, \|M_{N(n,s)+1}^{(n)} - M_t^{(n)}\|\}$. As a consequence,

$$\mathbb{P}\left(\sup_{s\in[t,t+\delta]} \|M_s^{(n)} - M_t^{(n)}\| \geq \varepsilon\right) \leq \mathbb{P}\left(\sup_{N(n,t)+1\leq k\leq N(n,t+\delta)+1} \|M_{\Gamma_k}^{(n)} - M_t^{(n)}\| \geq \varepsilon\right)$$

Let $p > 2$ and applying the Doob inequality, the assumption of the lemma leads to :

$$\mathbb{P}\left(\sup_{s\in[t,t+\delta]} \|M_s^{(n)} - M_t^{(n)}\| \geq \varepsilon\right) \leq \varepsilon^{-p}\mathbb{E}\left[\|M_{N(n,t+\delta)+1}^{(n)} - M_t^{(n)}\|^p\right]$$

and the Minkowski inequality yields :

$$\mathbb{P}\left(\sup_{s\in[t,t+\delta]} \|M_s^{(n)} - M_t^{(n)}\| \geq \varepsilon\right) \leq \varepsilon^{-p} \sum_{k=N(n,t)+1}^{N(n,t+\delta)+1} \gamma_k^{\frac{p}{2}} \mathbb{E}\left[\|\Delta M_k\|^p\right].$$

Under the assumptions of the lemma, $\mathbb{E}[[\|\Delta M_k\|^p] \leq C$. Furthermore, we can use the rough upper bound :

$$\sum_{k=N(n,t)+1}^{N(n,t+\delta)+1} \gamma_k^{\frac{p}{2}} \leq \gamma_n^{\frac{p}{2}-1} \sum_{k=N(n,t)+1}^{N(n,t+\delta)+1} \gamma_k.$$

Now, we can choose n_0 such that $\gamma_{n_0}^{p/2} \leq \epsilon^p\eta$ and we then obtain

$$\forall n \geq n_0 \qquad \mathbb{P}\left(\sup_{s\in[t,t+\delta]} \|M_s^{(n)} - M_t^{(n)}\| \geq \varepsilon\right) \leq \eta\delta.$$

This ends the proof of the tightness of $(M^{(n)})_{n\geq 1}$. ◇

The process $\bar{X}^{(n)}$ being the sum of tight processes, we obtain the conclusion of the proof □

Theorem 4.3.1 *Assume that f satisfies $(\mathbf{H}_{SC}(\mu, L))$ and that $\gamma_n = \gamma_1 n^{-\alpha}$ with $\alpha \in (0,1)$ or $\alpha = 1$ with $\gamma_1\mu > 1/2$. If $(\mathbf{H}_{\boldsymbol{\sigma},\boldsymbol{p}})$ holds, then $(\bar{X}^{(n)})_{n\geq 1}$ is tight.*

Proof : To prove this result, it is enough to check that the assumptions of Lemma 4.3.1 are satisfied. First, one remarks that our assumptions imply the convergence rates detailed in Section 4.1. Hence, we have the polynomial decay property :

$$\mathbb{E}[\|X_n - x^\star\|^2] \leq C\gamma_n,$$

because $\alpha < 1$ or $\alpha = 1$ and $\gamma_1\mu > 1/2$. . As a consequence, $\sup_{k\geq 1} \mathbb{E}[\|\check{X}_n\|^2] < +\infty$.

On the other hand, since we have assumed that $(\mathbf{H}_{\boldsymbol{\sigma},\boldsymbol{p}})$ holds and that $f(x)$ is bounded by $C(1 + \|x\|^2)$ for a suitable large C, we can derive that $\sup_n \mathbb{E}[f^p(X_n)] < +\infty$ and $(\mathbf{H}_{\sigma,\mathbf{p}})$ leads to :

$$\sup_{n\geq 1} \mathbb{E}[\|\Delta M_n\|^p] \lesssim \sup_n \mathbb{E}[f^p(X_n)] < +\infty.$$

Hence, we can apply Lemma 4.3.1 and obtain the tightness of $(\bar{X}^{(n)})_{n\geq 1}$. □

4.4 Central Limit Theorem

4.4.1 Main result

Below, we will show the following result.

Theorem 4.4.1 *Assume* $(\mathbf{H}_{SC}(\mu, L))$ *holds and that* $\alpha \in (0,1]$ *and if* $\alpha = 1$ *then* $\gamma\alpha > 1$. *Assume that* $(\mathbf{H}_{\boldsymbol{\sigma},\boldsymbol{p}})$ *holds with* $p > 2$ *when* $\alpha < 1$ *and* $p = \infty$ *when* $\alpha = 1$. *Finally, suppose that the following condition is fulfilled :*

$$\mathbb{E}\left[(\Delta M_{n+1})(\Delta M_{n+1})^t|\mathcal{F}_{n-1}\right] \xrightarrow{n\to+\infty} \Sigma^2 \quad \text{in probability} \tag{4.6}$$

where Σ^2 *is a symmetric positive* $p \times p$*-matrix. Then,*

(i) The normalized algorithm $\left(\frac{\check{X}_n}{\sqrt{\gamma_n}}\right)_n$ *converges in law to a centered Gaussian distribution* $\mu_\infty^{(\alpha)}$, *which is the invariant distribution of the (linear) diffusion with infinitesimal generator* $\mathcal{L}$ *defined on* $\mathcal{C}^2$*-functions by :*

$$\mathcal{L}g(z) = \left\langle \nabla g(z), \left(\frac{1}{2\gamma}1_{\{\alpha=1\}}I_{2d} - D^2 f(x^\star)\right) z \right\rangle + \frac{1}{2}\mathrm{Tr}(\Sigma^T D^2 g(z)\Sigma)$$

(ii) In the simple situation where $\Sigma^2 = \sigma_0^2 I_p$ *(*$\sigma_0 > 0$*) and* $\alpha < 1$. *In this case, the covariance of* $\mu_\infty^{(\alpha)}$ *is given by*

$$\frac{\sigma_0^2}{2}\{D^2 f(x^\star)\}^{-1}$$

In particular,

$$\frac{\check{X}_n}{\sqrt{\gamma_n}} \Longrightarrow \mathcal{N}(0, \frac{\sigma_0^2}{2}\{D^2 f(x^\star)\}^{-1}).$$

4.4.2 Identification of the limit

Starting from our compactness result above, we now characterize the potential weak limits of $(\bar{X}^{(n)})_{n\geq 1}$. This step is strongly based on the following lemma.

Lemma 4.4.1 *Suppose that the assumptions of Lemma 4.3.1 hold with* α *leq*1 *with* $\gamma_1\mu > 1/2$ *if* $\alpha = 1$ *and that :*

$$\mathbb{E}[\Delta M_n(\Delta M_n)^t|\mathcal{F}_{n-1}] \xrightarrow{n\to+\infty} \Sigma^2 \quad \text{in probability,}$$

where Σ^2 *is a positive symmetric* $d\times d$*-matrix. Then, for every* C^2*-function* $g : \mathbb{R}^p \to \mathbb{R}$, *compactly supported with Lipschitz continuous second derivatives, we have :*

$$\mathbb{E}(g(\check{X}_{n+1}) - g(\check{X}_n)|\mathcal{F}_n) = \gamma_{n+1}\mathcal{L}g(\check{X}_n) + R_n^g$$

where $\gamma_{n+1}^{-1}R_n^g \to 0$ *in* L^1 *and* $\mathcal{L}$ *is the infinitesimal generator defined by*

$$\forall\phi \in \mathcal{C}^2(\mathbb{R}^p) \qquad \mathcal{L}\phi(x) = \langle\left(\mathbf{1}_{\alpha=1}\frac{1}{2\gamma_1}I_p - D^2 f(x^\star)\right) x, \partial_x\phi\rangle + \frac{1}{2}\Sigma^t D^2\phi(x)\Sigma$$

Remark 4.4.1 *We recall some basic facts on Markov generator and stochastic differential equations. If we denote by* $\mathcal{L}$ *the infinitesimal generator of the following stochastic differential equation :*

$$dU_t = -AU_t dt + \Sigma dB_t,$$

where : A is a symmetric positive definite matrix and Σ^2 is a covariance matrix, then $(U_t)_{t\geq 0}$ lies in the family of Ornstein-Uhlenbeck processes. On the one hand, the drift and diffusion coefficients being respectively linear and constant, $(U_t)_{t\geq 0}$ is a Gaussian diffusion ; on the other hand, since A has negative eigenvalues, $(U_t)_{t\geq 0}$ is ergodic. This process is positive recurrent, elliptic and admits a unique centered Gaussian invariant measure whose variance will be characterised below...

Proof : C will denote an absolute constant whose value may change from line to line, for the sake of convenience. We use a Taylor expansion between $\check{X}_n$ and $\check{X}_{n+1}$ and obtain that a ξ_n exists in $[0,1]$ such that :

$$\begin{aligned} g(\check{X}_{n+1}) - g(\check{X}_n) &= \langle \nabla g(\check{X}_n), \check{X}_{n+1} - \check{X}_n \rangle + \frac{1}{2}(\check{X}_{n+1} - \check{X}_n)^T D^2 g(\check{X}_n)(\check{X}_{n+1} - \check{X}_n) \qquad (4.7) \\ &+ \underbrace{\frac{1}{2}(\check{X}_{n+1} - \check{X}_n)^T (D^2 g(\xi_n \check{X}_n + (1-\xi_n)\check{X}_{n+1}) - D^2 g(\check{X}_n))(\check{X}_{n+1} - \check{X}_n)}_{R^{(1)}_{n+1}} . \end{aligned}$$

We first deal with the remainder term $R^{(1)}_{n+1}$ and observe that $(\bar{C}_n)$ introduced in Equation (4.3) is uniformly bounded so that a constant C exists such that $\|b_n(z)\| \leq C\|z\|$. We thus conclude that :

$$\|\check{X}_{n+1} - \check{X}_n\| \leq C\left(\gamma_{n+1}\|\check{X}_n\| + \sqrt{\gamma_{n+1}}\|\Delta M_{n+1}\|\right).$$

Using $(\mathbf{H}_{\boldsymbol{\sigma},\boldsymbol{p}})$, we deduce that for any $\bar{p} \leq p$,

$$\mathbb{E}\left[\|\check{X}_{n+1} - \check{X}_n\|^{\bar{p}}\right] \leq C\gamma_{n+1}^{\frac{\bar{p}}{2}}. \qquad (4.8)$$

Since $D^2 g$ is Lipschitz continuous and compactly supported, $D^2 g$ is also ε-Hölder for all $\varepsilon \in (0,1]$. We choose ε such that $2+\varepsilon \leq p$ and obtain :

$$\mathbb{E}\left[|R^{(1)}_{n+1}|\right] \leq C\mathbb{E}\left[\|\check{X}_{n+1} - \check{X}_n\|^{2+\varepsilon}\right] \leq C\gamma_{n+1}^{1+\frac{\varepsilon}{2}}.$$

We deduce that $\gamma_{n+1}^{-1} R^{(1)}_{n+1} \to 0$ in L^1. ◇

Second, we can express Equation (4.4) when $\gamma_n = \gamma n^{-\alpha}$ with $\alpha \in (0,1]$ in the following form :

$$\epsilon_n := \frac{1}{\gamma_{n+1}}\left(\sqrt{\frac{\gamma_n}{\gamma_{n+1}}} - 1\right) - \frac{1}{2\gamma}1_{\{\beta=1\}} = o(1).$$

Then, given that $D^2 f$ is Lipschitz (and that $x^\star = 0$), it follows that :

$$\forall z \in \mathbb{R}^p \qquad \left\| b_n(z) - \left(\frac{1}{2\gamma}1_{\{\alpha=1\}} I_p - D^2 f(x^\star)\right) z \right\| \leq (\varepsilon_n + \|\check{X}_n\|)\|z\|$$

where $(\varepsilon_n)_{n\geq 1}$ is a deterministic sequence such that $\lim_{n\to+\infty} \varepsilon_n = 0$.

We may apply the convergence rates described in Section 4.1 and observe that $\sup_n \mathbb{E}[\|X_n\|^2] \lesssim \gamma_n$, meaning that $\sup_n \mathbb{E}[\|\check{X}_n\|^2] < +\infty$. We deduce that :

$$\mathbb{E}[\langle \nabla g(\check{X}_n), (\check{X}_{n+1} - \check{X}_n)\rangle | \mathcal{F}_n] = \gamma_{n+1}\langle \nabla g(\check{X}_n), \left(\frac{1}{2\gamma}1_{\{\alpha=1\}} I_p - D^2 f(x^\star)\right)\check{X}_n\rangle + R^{(2)}_n$$

where $\gamma_{n+1}^{-1} R^{(2)}_n \to 0$ in L^1 as $n \to +\infty$. Let us now consider the second term of the right-hand side of Equation (4.7). We have :

$$\mathbb{E}[(\check{X}_{n+1} - \check{X}_n)^T D^2 g(\check{X}_n)(\check{X}_{n+1} - \check{X}_n)|\mathcal{F}_n] = \gamma_{n+1}\sum_{i,j} D^2_{x_i x_j} g(\check{X}_n)\mathbb{E}[\Delta M^i_{n+1}\Delta M^j_{n+1}|\mathcal{F}_n] + R^{(3)}_n$$

where

$$|\gamma_{n+1}^{-1} R_n^{(3)}| \leq C\gamma_{n+1}\|\check{X}_n\|^2 \xrightarrow{n\to+\infty} 0 \quad \text{in } L^1$$

under the assumptions of the lemma. To conclude the proof, it remains to note that under the assumptions of the lemma for any i and j, $(\mathbb{E}[\Delta M_{n+1}^i \Delta M_{n+1}^j|\mathcal{F}_n])_{n\geq 1}$ is a uniformly integrable sequence that satisfies :

$$\mathbb{E}[\Delta M_{n+1}^i \Delta M_{n+1}^j|\mathcal{F}_n] = \Sigma_{i,j} \quad \text{in probability.}$$

Thus, the convergence also holds in L^1. The conclusion of the lemma easily follows from the boundedness of $D^2 g$. $\diamond\square$

We are now able to prove Theorem 4.4.1 :

Proof of Theorem 4.4.1, (i) **-** $\alpha < 1$: Note that under the assumptions of Theorem 4.4.1, we can apply Lemma 4.3.1 and Lemma 4.4.1 and obtain that the sequence of processes $(\bar{X}^{(n)})_{n\geq 1}$ is tight. The rest of the proof is then divided into two steps. In the first one, we prove that every weak limit of $(\bar{X}^{(n)})_{n\geq 1}$ is a solution of the martingale problem $(\mathcal{L}, \mathcal{C})$ where $\mathcal{C}$ denotes the class of $\mathcal{C}^2$-functions with compact support and Lipschitz-continuous second derivatives. Before going further, let us recall that, owing to the Lipschitz continuity of the coefficients, this martingale problem is well-posed, *i.e.*, that existence and uniqueness hold for the weak solution starting from a given initial distribution μ (see, *e.g.*, Ethier-Kurtz or Stroock-Varadhan).

In a second step, we prove the uniqueness of the invariant distribution related to the operator $\mathcal{L}$ and the convergence in distribution to this invariant measure. We end this proof by showing that $(\bar{X}^{(n)})$ converges to this invariant distribution, so that the sequence $(\bar{X}^{(n)})_{n\geq 1}$ converges to a stationary solution of the previously introduced martingale problem. We will characterize this invariant (Gaussian) distribution in the next paragraph.

Step 1 : Let g belong to $\mathcal{C}$ and let $(\mathcal{F}_t^{(n)})_{t\geq 0}$ be the natural filtration of $\bar{X}^{(n)}$. To prove that any weak limit of $(\bar{X}^{(n)})_{n\geq 1}$ solves the martingale problem $(\mathcal{L}, \mathcal{C})$, it is enough to show that :

$$\forall t \geq 0, \quad g(\bar{X}_t^{(n)}) - g(\bar{X}_0^{(n)}) - \int_0^t \mathcal{L}g(\bar{X}_s^{(n)})ds = \mathcal{M}_t^{(n,g)} + \mathcal{R}_t^{(n,g)}$$

where $(\mathcal{M}_t^{(n,g)})_{t\geq 0}$ is an $(\mathcal{F}_t^{(n)})$-adapted martingale and $\mathcal{R}_t^{(n,g)} \to 0$ in probability for any $t \geq 0$. We set :

$$\mathcal{M}_t^{(n,g)} = \sum_{k=n+1}^{N(n,t)} g(\bar{X}_{k+1}) - g(\bar{X}_k) - \mathbb{E}[g(\check{X}_{k+1}) - g(\check{X}_k)|\mathcal{F}_{k-1}].$$

By construction, $(\mathcal{M}_t^{(n,g)})_{t\geq 0}$ is an $(\mathcal{F}_t^{(n)})$-adapted martingale (given that $\mathcal{F}_s^{(n)} = \mathcal{F}_{\underline{s}_n}^{(n)}$) and :

$$\mathcal{R}_t^{(n,g)} = g(\bar{X}_t^{(n)}) - g(\bar{X}_{\underline{t}_n}^{(n)}) - \int_{\underline{t}_n}^t \mathcal{L}g(\bar{X}_s^{(n)})ds + \int_0^{\underline{t}_n} \left(\mathcal{L}g(\bar{X}_{\underline{s}_n}^{(n)}) - \mathcal{L}g(\bar{X}_s^{(n)})\right) ds + \sum_{k=n}^{N(n,t)-1} R_k^g$$

where $(R_k^g)_{k\geq 1}$ has been defined in Lemma 4.4.1. Using an argument similar to Equation (4.8), we can check that for any $t \geq 0$:

$$\sup_{s\leq t} \mathbb{E}[\|\bar{X}_s^{(n)} - \bar{X}_{\underline{s}_n}^{(n)}\|^2] \leq C\sqrt{\gamma}_n.$$

This inequality combined with the Lipschitz continuity of g and its derivatives implies that the first three terms tend to 0 when $n \to +\infty$. Now, concerning the last one, the previous lemma yields :

$$\mathbb{E}\left[\left|\sum_{k=n}^{N(n,t)-1} R_k^g\right|\right] \leq Ct \sup_{k\geq n} \mathbb{E}\left[\left|\gamma_k^{-1} R_k^g\right|\right] \xrightarrow{n\to+\infty} 0.$$

$\diamond$

Step 2 : First, let us prove that uniqueness holds for the invariant distribution related to $\mathcal{L}$. We denote it by $\mu_\infty^{(\beta)}$ below. In this simple setting where the coefficients are linear, we could use the fact that the process, which is solution to the martingale problem, is Gaussian so that any invariant distribution is so. Uniqueness could then be deduced through the characterization of the mean and the variance through the relationship $\int \mathcal{L}f(x)\mu_\infty^{(\alpha)}(dx) = 0$ (see next subsection for such an approach). However, at this stage, we prefer to use a more general strategy related to the ellipticity of $\mathcal{L}$: the matrix Σ^2 is positive definite and symmetric. Therefore, the semi-group is elliptic and the process irreducible. Of course, irreducibility implies uniqueness of the potential invariant distribution.

Second, it can be shown that the Orstein-Uhlenbeck semi-group is ergodic because the drift matrix is positive definite and satisfies :

$$\sup_{z\in K} |P_t g(z) - \mu_\infty^{(\alpha)} g| \longrightarrow 0 \qquad \text{as} \qquad t \longrightarrow +\infty,$$

where K is an arbitrary compact set of $\mathbb{R}^p$.

Step 3 : Let $(\check{X}_{n_k})_{k\geq 1}$ be a (weakly) convergent subsequence of $(\check{X}_n)_{n\geq 1}$ to a probability ν. We have to prove that $\nu = \mu_\infty^{(\alpha)}$. To do this, we take advantage of the "shifted" construction of the sequence $(\bar{X}^{(n)})_{n\in\mathbb{N}}$. More precisely, as a result of construction, for any positive T, a sequence $(\psi(n_k,T))_{k\geq 1}$ exists such that :

$$N(T, \psi(n_k,T)) = n_k.$$

In other words,

$$\bar{X}^{(\psi(n_k,T))}_{\underline{T}_{\psi(n_k,T)}} = \bar{X}_{n_k}.$$

At the price of a potential extraction, $(\bar{X}^{(\psi(n_k,T))})_{k\geq 1}$ is convergent to a continuous process, which is denoted by $X^{\infty,T}$ below. Given that $\bar{X}^{(n)}_T - \bar{X}^{(n)}_{\underline{T}_n}$ tends to 0 as $n \to +\infty$ in probability, it follows that $X^{\infty,T}_T$ has distribution ν. However, according to Step 1, $X^{\infty,T}$ is also a solution to the martingale problem $(\mathcal{L},\mathcal{C})$ so that for any Lipschitz continuous function g,

$$\mathbb{E}[g(X^{\infty,T}_T)] - \mu_\infty^{(\alpha)}(g) = \int_{\mathbb{R}^p} \Big(P_T g(z) - \mu_\infty^{(\alpha)}(g)\Big)\, \mathbb{P}_{X_0^{\infty,T}}(dz).$$

Denote by $\mathcal{P}$, the set of weak limits of $(\bar{X}_n)_{n\geq 1}$. $\mathcal{P}$ is tight and as a result of construction, $X_0^{\infty,T}$ belongs to $\mathcal{P}$. Thus, for any $\varepsilon > 0$, a compact set K_ε exists such that for any $T > 0$,

$$\left| \int_{K_\varepsilon^c} \Big(P_T g(z) - \mu_\infty^{(\alpha)}(g)\Big)\, \mathbb{P}_{X_0^{\infty,T}}(dz) \right| \leq 2\|g\|_\infty \sup_{\mu\in\mathcal{P}} \mu(K_\varepsilon^c) \leq 2\|g\|_\infty \varepsilon.$$

On the other hand,

$$\left| \int_{K_\varepsilon} \Big(P_T g(z) - \mu_\infty^{(\alpha)}(g)\Big)\, \mathbb{P}_{X_0^{\infty,T}}(dz) \right| \leq \sup_{z\in K_\varepsilon} |P_T g(z) - \mu_\infty^{(\alpha)}(g)|$$

and it follows from Step 2 that the right-hand member tends to 0 as $T \to +\infty$. From this, we can therefore conclude that for any bounded Lipschitz-continuous function g, a large enough T exists such that :

$$\Big| \mathbb{E}[g(Z^{\infty,T}_T)] - \mu_\infty^{(\alpha)}(g) \Big| \leq C_g \varepsilon.$$

Since $\mathbb{E}[g(Z^{\infty,T}_T)] = \nu(g)$, it follows that $\nu(g) = \mu_\infty^{(\alpha)}(g)$. Finally, the set $\mathcal{P}$ is reduced to a single element $\mathcal{P} = \{\mu_\infty^{(\beta)}\}$, and the whole sequence $(\bar{X}_n)_{n\geq 1}$ converges to $\mu_\infty^{(\beta)}$.

Before ending this section, let us note that $\mu_\infty^{(\alpha)}$ is a Gaussian centered distribution : this is a simple consequence of Remark 4.4.1. We therefore leave this point to the reader. $\diamond\square$

4.4.3 Identifiying the limit variance

We end this section on the analysis of the rescaled algorithm with some considerations on the invariant measure $\mu_\infty^{(\alpha)}$ involved in Theorem 4.4.1. As shown in the above paragraph, this invariant measure describes the exact asymptotic variance of the initial algorithm. We now focus on its characterization *i.e.*, on the proof of Theorem 4.4.1(ii). In particular, to ease the presentation, we assume that the covariance matrix Σ^2 related to $(\Delta M_{n+1})_{n\geq 1}$ is proportional to the identity matrix :

$$\lim_{n\longrightarrow+\infty} \mathbb{E}\left[\Delta M_{n+1}\left(\Delta M_{n+1}\right)^t|\mathcal{F}_n\right] = \sigma_0^2 I_d \quad \text{in probability.} \tag{4.9}$$

We also assume that $\gamma_n = \gamma n^{-\alpha}$ with $\alpha < 1$. Then, (i) of Theorem 4.4.1 states that $(\bar{X}_n)_{n\geq 1}$ weakly converges toward a diffusion process, whose generator $\mathcal{L}$ is the one of an Ornstein-Uhlenbeck process. Assumption Equation (4.9) leads to a simpler expression :

$$\mathcal{L}(\phi)(x) = -\langle D^2 f(x^\star)x, \nabla_x\phi\rangle + \frac{{\sigma_0}^2}{2}\Delta_x\phi. \tag{4.10}$$

A particular feature of Equation Equation (4.10) when $\gamma_n = \gamma n^{-\alpha}$ is that $\mathcal{L}$ does not depend on α nor γ. The invariant measure $\mu_\infty^{(\alpha)}$ is a multivariate Gaussian distribution that may be well described in the basis given by the eigenvectors of the Hessian $D^2(f)(x^\star)$. The reduction to p unidimensional system makes it possible to use the spectral decomposition of $D^2(f)(x^\star) = P^{-1}\Lambda P$ where P is an orthonormal matrix and Λ a diagonal matrix with positive eigenvalues. The process $\tilde{X}^{(n)}$ is therefore centered and Gaussianly distributed asymptotically. This process is associated with p blockwise independent Ornstein-Uhlenbeck processes. Given λ the eigenvalue associated to any eigenvector, the generator on this coordinate is now

$$\check{\mathcal{L}}(\phi)(\check{x}) = -\lambda\langle\check{x}, \nabla_{\check{x}}\phi\rangle + \frac{{\sigma_0}^2}{2}\Delta_{\check{x}}\phi,$$

where we used $\mathrm{Tr}\left(P^t D^2_{\check{x}} P\right) = \mathrm{Tr}\left(D^2_{\check{x}} P P^t\right) = \mathrm{Tr}\left(D^2_{\check{x}}\right)$ in the last line because $P^t P = I_d$. If we denote $\check{\mu}_\infty^{(\alpha)}$ the associated invariant gaussian measure, the tensor structure of $\check{\mathcal{L}}$ and the relationship $\displaystyle\int \check{\mathcal{L}}(\phi)d\check{\mu}_\infty^{(\alpha)} = 0$ for some well chosen functions ϕ, we can identify the covariance matrix. Denote i any integer in $\{1,\ldots,p\}$ (that points a coordinate on an eigenvector). We choose $\phi(\check{x}) = \frac{\{\check{x}^{(i)}\}^2}{2}$ and obtain that $\check{\mathcal{L}}\left(\frac{\{\check{x}^{(i)}\}^2}{2}\right)(\check{x}) = -\lambda_i\{\check{x}^{(i)}\}^2 + \frac{\sigma_0^2}{2}$. It then implies that

$$\mathbb{E}_{\check{x}\sim\check{\mu}_\infty^{(\alpha)}}[\{\check{x}^{(i)}\}^2] = \frac{\sigma_0^2}{2\lambda_i}. \tag{4.11}$$

Picking now $\phi(\check{x}) = \check{x}^{(i)}\check{x}^{(j)}$, we obtain $\check{\mathcal{L}}\left(\check{x}^{(i)}\check{x}^{(j)}\right)(\check{x},\check{y}) = -(\lambda_i+\lambda_j)\check{x}^{(i)}\check{x}^{(j)}$ so that

$$\mathbb{E}_{(\check{x},\check{y})\sim\check{\mu}_\infty^{(\alpha)}}[\check{x}^{(i)}\check{x}^{(j)}] = 0. \tag{4.12}$$

We can sum-up formulae Equation (4.11)-Equation (4.12) in $\check{\mu}_\infty^{(\alpha)} = \mathcal{N}\left(0, \frac{\sigma_0^2}{2}\Lambda^{-1}\right)$. □

Chapitre 5

Stabilisation of Markov processes

5.1 Semi-group, Markov process, infinitesimal generator

We denote by E the state space (Polish space is a minimal requirement). For the sake of simplicity, we assume that $E = \mathbb{R}^d$ and $\mathcal{E}$ is the set of Borelians. We recall the definition of Markov kernels :

Definition 5.1.1 *Une application de E dans $\mathcal{P}(E)$ de la forme $x \mapsto P(x,.)$ est appelée une probabilité de transition sur E si*
(i) Pour tout $A \in \mathcal{E}$, $x \mapsto P(x,A)$ est $\mathcal{E}$-mesurable.
(ii) Pour tout $x \in \mathcal{E}$, $A \mapsto P(x,A)$ is a probability on E.

On a alors la définition suivante un semi-groupe $(P_t)_{t\geq 0}$:

Definition 5.1.2 (Semi-groupe) *On appelle semigroupe de transitions sur E que l'on note $(P_t)_{t\geq 0}$ une famille de noyaux de transitions dépendant d'un paramètre $t \in \mathbb{R}_+$, vérifiant*

$$\forall (t,s) \in \mathbb{R}_+^2 \qquad P_0 = Id \qquad et \qquad P_t \circ P_s = P_{t+s}.$$

(Autrement dit, pour toute fonction f $\mathcal{E}$-mesurable et bornée,

$$P_{t+s}f = P_t(P_s f) \quad \text{où} \quad P_t f(x) = \int_E f(y) P_t(x,dy).)$$

Notons $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ un espace de probabilité filtré. *

Definition 5.1.3 *Un processus $(X_t)_{t\geq 0}$ défini sur $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ et à valeurs dans E est appelé un processus de Markov s'il existe un semi-groupe de transitions $(P_t)_{t\geq 0}$ tel que pour tout $x \in E$, pour tous $s,t \geq 0$, pour toute fonction mesurable positive ou bornée,*

$$\mathbb{E}_x[f(X_{t+s})|\mathcal{F}_t] = P_s f(X_t).$$

On ne détaillera pas ici la construction d'un tel processus. On rappelle cependant que si E est polonais et si (P_t) est de Feller au sens suivant :

1. Pour toute fonction continue $E :\to \mathbb{R}$ nulle à l'infini (pour le compactifié d'Alexandrov) , pour tout $t \geq 0$ $x \mapsto P_t f(x)$ est continue et nulle à l'infini (pour le compactifié d'Alexandrov) :
2. (P_t) est stochastiquement continu, *i.e.* pour toute fonction continue $f : E \mapsto \mathbb{R}$, pour tout $x \in \mathbb{R}^d$, $P_t f(x) \to x$ lorsque t tend vers 0,

alors (X_t) admet une version càdlàg, *i.e.* continue à droite et admettant des limites à gauche et que la propriété de Markov forte est valide.

Remark 5.1.1 *Dans la littérature, un semi-groupe est souvent dit féllerien dès que f continue bornée implique $P_t f$ continue bornée pour tout t.*

Dans la suite, le processus (ou le semi-groupe) sera supposé de Feller.

Générateur infinitésimal :

Definition 5.1.4 *Soit (X_t) un processus de Feller. Considérons $f \in \mathcal{C}_0$ (espace des fonctions continues bornées nulles à l'infini) et Supposons que*

$$\lim_{t\to 0^+} \frac{1}{t}(P_t f - f) =: \mathcal{L}f$$

existe. Alors, on dit que $f \in \mathcal{D}(\mathcal{L})$ si de plus, $\mathcal{L}f$ appartient à $\mathcal{C}_0$. L'opérateur $\mathcal{L} : \mathcal{D}(\mathcal{L}) \to \mathcal{C}_0$ est alors appelé le générateur infinitésimal du semi-groupe $(P_t)_{t\geq 0}$.

Le générateur infinitésimal permet donc ainsi de décrire l'évolution infinitésimale du processus. On a le résultat suivant :

Proposition 5.1.1 *Soit $f \in \mathcal{D}(\mathcal{L})$. Supposons de plus que*

$$(\mathbf{H_f}) : \qquad \sup_{t\leq 1, x\in E} \frac{|P_t f(x) - f(x)|}{t} < +\infty.$$

Alors,

1. *Pour tout $t \geq 0$, $P_t f \in \mathcal{D}(\mathcal{L})$.*
2. *La fonction $t \mapsto P_t f$ est dérivable à droite et pour tout $t \geq 0$,*

$$\frac{d^+}{dt} P_t f = \mathcal{L} P_t f = P_t(\mathcal{L}f).$$

3. *Pour tout $t \geq 0$,*

$$\forall x \in E, \quad P_t f(x) = f(x) + \int_0^t \mathcal{L} P_s f ds.$$

Remark 5.1.2 *Le dernier point fournit parfois la définition du générateur dans la littérature.*

<u>Preuve</u> : Pour tout $t \geq 0$,

$$\lim_{s\to 0^+} \frac{P_s(P_t f) - f}{s} = \lim_{s\to 0^+} P_t \left(\frac{P_s f - f}{s} \right).$$

Par la définition de $\mathcal{L}f$ et l'hypothèse $(\mathbf{H_f})$, on déduit alors du théorème de Lebesgue que

$$\lim_{s\to 0^+} P_t \left(\frac{P_s f - f}{s} \right) = P_t \left(\lim_{s\to 0^+} \frac{P_s f - f}{s} \right) = P_t \mathcal{L} f.$$

Ceci nous permet d'en déduire les deux premiers points. En intégrant et en prenant la valeur en 0, on en déduit le dernier. □

Remark 5.1.3 *On peut aussi montrer que $\mathcal{D}(\mathcal{L})$ est dense dans $\mathcal{C}_0$ (pour la topologie de la convergence simple). En pratique, on travaillera généralement sur un sous-espace de $\mathcal{D}(\mathcal{L})$ qui est lui-aussi dense dans $\mathcal{C}_0$.*

On énonce enfin une proposition importante à la fois pour la pratique et pour la compréhension du lien fort qui existe entre processus de Markov et martingales :

Proposition 5.1.2 *Supposons que $f \in \mathcal{D}(\mathcal{L})$ et que $(\mathbf{H_f})$ soit satisfaite. Alors, (M_t^f) définie par*

$$\forall t \geq 0, \quad M_t^f = f(X_t) - f(X_0) - \int_0^t \mathcal{L}f(X_s)ds$$

est une $(\mathcal{F}_t, \mathbb{P}_\mu)$-martingale (où $\mathcal{F}_t = \sigma(X_s, s \leq t)$ et $\mathbb{P}_\mu$ est la loi du processus $(X_t)_{t\geq 0}$ de loi initiale μ.

Remark 5.1.4 *On dit aussi que $\mathbb{P}_\mu$ est solution du problème de martingale $(\mathcal{A}, \mathcal{D}(\mathcal{A}))$. Il y a donc une sorte de réciproque à ce résultat : si pour une classe de fonctions f suffisamment grande (pour être caractérisante), il existe une et une seule loi $\mathbb{P}_\mu$ sous laquelle (M_t^f) est une $(\mathcal{F}_t, \mathbb{P}_\mu)$-martingale (pour donner un sens à ce qui précède, il faut se placer sur l'espace canonique associé au processus, mais on restera vague sur ce point). Ce type d'outil peut être particulièrement puissant pour déterminer la loi d'un processus limite par exemple. On renvoit au livre d'Ethier et Kurtz pour plus de détails sur ce vaste sujet.*

Remark 5.1.5 *Lorsque f n'est pas bornée, $\mathcal{L}f$ peut conserver un sens. Dans ce cas, (M_t^f) sera alors a priori une martingale locale.*

<u>Preuve</u> : f et $\mathcal{L}f$ appartiennent à $\mathcal{C}_0$ de sorte que $M_t f$ est intégrable pour tout $t \geq 0$. De plus, pour tous $0 \leq s < t$,

$$\mathbb{E}[M_t^f - M_s^f|\mathcal{F}_s] = \mathbb{E}[f(X_t) - f(X_s) - \int_s^t \mathcal{L}f(X_u)du|\mathcal{F}_s] = P_{t-s}f(X_s) - f(X_s) - \int_0^{t-s} P_v\mathcal{L}f(X_s)dv$$

où dans la dernière ligne, on a appliqué Fubini (en utilisant que $\mathcal{L}f$ est bornée). Ainsi, d'après la proposition précédente, on en déduit que cette quantité est égale à 0. □ **Exemples :**

— Soit (X_t) processus à valeurs dans $\mathbb{R}^d$ solution de $dX_t = b(X_t)dt + \sigma(X_t)dB_t$ où b et σ sont Lipschitziennes. Alors, par la formule d'Itô, on peut vérifier que pour toute fonction $f \in \mathcal{C}_K^2$ ($\mathcal{C}^2$ à support compact),

$$\mathbb{E}[f(X_t^x)] = f(x) + \int_0^t \mathbb{E}[\mathcal{L}f(X_s^x)]ds$$

avec

$$\mathcal{L}f = \langle \nabla f, b\rangle + \frac{1}{2}\sum_{i,j}(\sigma\sigma^*)_{i,j}\partial^2_{x_i,x_j}f.$$

— Soit $(N_t)_{t\geq 0}$ un processus de Poisson composé d'intensité λ et de loi de saut μ. Alors, $\mathcal{D}(\mathcal{L} = \mathcal{C}_0$ et

$$\mathcal{L}f(x) = \lambda \int (f(x+y) - f(x))\mu(dy).$$

5.2 Mesures invariantes : définition et existence

5.2.1 Définition, caractérisation

Definition 5.2.1 (Mesures invariantes pour $(P_t)_{t\geq 0}$) *μ est invariante pour le semi-groupe si pour toute fonction f mesurable (positive ou bornée), on a*

$$\int_{x\in E} P_t f(x)d\mu(x) = \int_{x\in E} f(x)d\mu(x).$$

Ce qui se réécrit en $\mu(P_t f) = \mu(f)$.

L'action en tout temps des transitions dans P_t laisse donc invariante μ au travers de l'intégrale sur les fonctions de E. L'invariance d'une loi μ peut aussi être caractérisée au travers du générateur.

Proposition 5.2.1 *Soit $(P_t)_{t\geq 0}$ un semi-groupe et $(\mathcal{L}, \mathcal{D}(\mathcal{L}))$ le générateur infinitésimal associé. Soit μ une probabilité sur E. Alors,*

$$\mu \text{ invariante pour } \mathcal{L} \iff \forall f \in \mathcal{D}(\mathcal{L}), \int \mathcal{L} f d\mu = 0.$$

Remark 5.2.1 *Pour prouver ce résultat, on va utiliser que $\mathcal{D}(\mathcal{L})$ est suffisamment gros. Ce point un peu délicat est laissé de côté.*

<u>Preuve</u> : On admet que lorsque le processus est de Feller, l'ensemble

$$B := \{f \in \mathcal{D}(\mathcal{L}),\ (\mathbf{H_f}) \text{ est satisfaite}\}$$

est dense dans $\mathcal{C}_0$ pour la topologie de la convergence simple. Ainsi, si $\mu(P_t f) = \mu(f)$ pour tout $f \in B$, alors $\mu(P_t f) = \mu(f)$ pour tout $f \in \mathcal{C}_0$ (en appliquant le théorème de Lebesgue). Il suffit donc de montrer que pour tout $f \in B$, $\mu(P_t f) = \mu(f)$. Comme pour tout $f \in B$, la formule de Dynkin et la commutation sont vraies, on a pour une telle fonction (en appliquant Fubini, ce qui est possible car par définition $\mathcal{L}f$ est bornée donc $\mathcal{L}P_s f = P_s \mathcal{L} f$ également) :

$$\mu(P_t f) = \mu(f) + \int_0^t \mu(\mathcal{L}P_s f) ds.$$

Or, $P_s f \in \mathcal{D}(\mathcal{L})$ de sorte que pour tout s, $\mu(\mathcal{L}P_s f) = 0$. On en déduit le résultat. □

L'existence de mesure stationnaire est issue d'arguments qui seront également valables pour les chaines de Markov à espace d'états plus que dénombrable. L'idée principale pour établir un tel résultat revient à montrer que le processus markovien passe la plupart de son temps dans un espace borné puis il convient d'utiliser un argument de type compacité pour prouver l'existence d'une telle mesure invariante.

Il y a alors deux façons de procéder :

— la première technique consiste à démontrer que les mesures d'occupation aléatoires d'une trajectoire, formées par

$$\mu_n = \frac{1}{t_n} \int_0^{t_n} \delta_{X_s} ds$$

forment un ensemble tendu de mesures, ce qui permet d'extraire une sous-suite convergente puis en déduire que la limite est stable par P_t. Ce genre d'arguments est employé par exemple dans le livre
[EK86] S. Ethier et T. Kurtz *Markov processes, characterisation and convergence* Wiley, New-York, 1986.

— La seconde méthode consiste à exhiber une mesure qui est invariante pour le processus markovien. La construction fait appel à la définition d'une chaîne de squelette pour le processus $(X_t)_{t\geq 0}$ et est utilisée dans
[K12] R. Khasminskii *Stochastic stability of Differential Equations*, Second Edition, Stochastic Modelling and Applied Porbability, Springer.
Comme cette construction est également fondamentale pour le problème des grandes déviations de mesure invariante (chapitre 4) et le recuit simulé, nous aborderons cette construction là.

Quelle que soit la méthode employée, il s'agit en général de trouver une façon d'estimer les temps de retour dans les espaces compacts. Ceci est en général facile dès lors qu'on a trouvé une « bonne » fonction de Lyapunov.

5.2.2 Existence de mesure invariante (cadre topologique)

Si on généralise les arguments développés dans le chapitre 1, alors on comprend que si E est compact (et polonais), alors comme $\mathcal{P}(E)$ est aussi compact (théorème de Prokhorov sur un espace polonais), on va pouvoir construire des mesures invariantes comme limites faibles de "mesure d'occupation moyennées" (voir plus bas, les familles (μ_t)). Mais lorsque $E = \mathbb{R}^d$ par exemple, alors il faut combler le manque de compacité de l'espace en s'appuyant sur la dynamique du processus. On va en fait demander au processus d'avoir une "propriété de rappel", *i.e.* d'avoir la propriété d'avoir tendance à revenir lorsque l'on s'éloigne trop de 0. On exprimera cette propriété à partir d'une fonction $V : \mathbb{R}^d \to \mathbb{R}$ continue, strictement positive et coercive, *i.e.* telle que $\lim_{|x|\to+\infty} V(x) = +\infty$ (Ici, $|\,.\,|$ désigne la norme euclidienne sur $\mathbb{R}^d$).

Avant d'énoncer le résultat, on rappelle quelques définitions et un critère de tension basé sur ce type de fonction. Dans ce qui suit, on supposera que $\mathcal{P}(E)$ est muni de la topologie de la convergence étroite : une suite $(\nu_n)_{n\geq 1}$ de $\mathcal{P}(E)$ converge vers $\nu \in \mathcal{P}(E)$ si pour toute fonction $f : E \to \mathbb{R}$ continue et bornée, $\nu_n(f) \to \nu(f)$ lorsque $n \to +\infty$.

Definition 5.2.2 *(et Théorème de Prokhorov) Soit E un espace topologique. Soit $\mathcal{M} \subset \mathcal{P}(E)$. On dit que $\mathcal{M}$ est tendu si pour tout $\varepsilon > 0$, il existe un compact K_ε tel que pour tout $\nu \in \mathcal{M}$, $\nu(K_\varepsilon^c) \leq \varepsilon$. Si E est polonais, alors $\mathcal{M}$ est tendu si et seulement si $\mathcal{M}$ est relativement compact. En particulier, toute suite (ν_n) de $\mathcal{M}$ admet une sous-suite convergente.*

Lorsque $E = \mathbb{R}^d$, on a le critère suivant :

Proposition 5.2.2 *Soit $\mathcal{M} \subset \mathcal{P}(\mathbb{R}^d)$. Alors, s'il existe une fonction continue $V : \mathbb{R}^d \to \mathbb{R}_+$ telle que $\lim_{|x|\to+\infty} V(x) = +\infty$ et telle que*

$$\sup_{\nu\in\mathcal{M}} \nu(V) < +\infty,$$

alors $\mathcal{M}$ est tendu.

<u>Preuve</u> : Notons $C = \sup_{\nu\in\mathcal{M}} \nu(V)$. Notons $K_\delta := \{x \in \mathbb{R}^d, V(x) \leq \delta\}$. Comme V est coercive, K_δ est compact. Or, par définition,

$$1_{K_\delta^c} \leq \frac{V(x)}{\delta} \Longrightarrow \nu(K_\delta^c) \leq \frac{\nu(V)}{\delta}.$$

Pour ε fixé, il suffit alors de poser $\delta = C/\varepsilon$. $\square$ On énonce maintenant un résultat d'existence de mesure invariante basé sur la fonction de Lyapounov.

Theorem 5.2.1 *Soit $(X_t)_{t\geq 0}$ un processus de Markov à valeurs dans $\mathbb{R}^d$ de semi-groupe $(P_t)_{t\geq 0}$ et de générateur infinitésimal $\mathcal{L}$. Supposons qu'il existe $V : \mathbb{R}^d \to \mathbb{R}_+$ continue et coercive (telle que $\lim_{|x|\to+\infty} V(x) = +\infty$) telle que*

$$(i) \quad \forall (t,x) \in \mathbb{R}_+ \times E, \quad P_t V(x) = V(x) + \int_0^t P_s(\mathcal{L}V)(x)ds \tag{5.1}$$

$$\tag{5.2}$$

$$(ii) \quad \mathcal{L}V \leq \beta \in \mathbb{R} \quad \text{et} \quad \limsup_{|x|\to+\infty} \mathcal{L}V(x) < 0. \tag{5.3}$$

Alors, $(X_t)_{t\geq 0}$ admet au moins une probabilité invariante.

Remark 5.2.2 *Comme V n'est pas dans le domaine, ceci signifie en particulier qu'on suppose que $\mathcal{L}V := \lim_{t\to 0^+} \frac{P_tV-V}{t}$ est bien définie. On suppose également ici que la formule de Dynkin est satisfaite pour V, ce qui pas non plus automatique a priori dans ce cadre. Cette hypothèse pourrait être affaiblie. L'utilisation de fonctions de Lyapounov peut aussi avoir un intérêt en temps court. Par exemple, pour une diffusion à coefficients localement lipschitziens (et non lipschitziens) mais tels qu'il existe une fonction $V : \mathbb{R}^d \to \mathbb{R}_+$ positive et coercive telle que $\mathcal{L}V \leq CV$, alors $\sup_{t\in[0,T]} \mathbb{E}[V(X_t)] < +\infty$. On peut alors en déduire l'existence et l'unicité des solutions dans ce cadre.*

<u>Preuve</u> : On suppose pour simplifier que $\limsup_{|x|\to+\infty} \mathcal{L}V(x) = -\infty$ et que $\mathcal{L}V$ est continue. On note $(\mu_t)_{t\geq 1}$ la famille de probabilités définie pour toute fonction f mesurable et bornée par :

$$\forall t \geq 1, \quad \mu_t(f) = \frac{1}{t}\int_0^t P_s f(x)ds.$$

où $x \in \mathbb{R}^d$. On veut construire une mesure invariante pour (P_t) comme valeur d'adhérence de cette famille lorsque $t \to +\infty$: on veut d'abord montrer que $(\mu_t)_{t\geq 1}$ est tendue en appliquant la proposition 5.2.2. On considère g définie par $g(x) = \beta - \mathcal{L}V(x)$. g est positive et coercive. De plus,

$$\forall t \geq 1, \quad \mu_t(g) = \beta - \frac{1}{t}\int_0^t P_s(\mathcal{L}V(x)ds \leq \beta + \frac{V(x) - P_tV(x)}{t}.$$

On en déduit facilement que $\sup_{t\geq 1} \mu_t(g) < +\infty$ ce qui implique que $(\mu_t)_{t\geq 1}$ admet des valeurs d'adhérence. Il reste à vérifier que celles-ci sont des probabilités invariantes. Notons $\mu_\infty := \lim_{n\to+\infty} \mu_{t_n}$ (au sens de la convergence étroite). Comme le semi-groupe est de Feller, pour toute fonction $\mathcal{C}_0$, P_Tf est $\mathcal{C}_0$ pour tout T. Ainsi, par définition de la convergence étroite, pour tout $f \in \mathcal{C}_0$,

$$\mu_\infty P_T(f) - \mu_\infty(f) = \lim_{n\to+\infty} \left(\mu_{t_n}(P_Tf) - \mu_{t_n}(f)\right).$$

Or, par un simple changement de variable

$$\begin{aligned}\mu_{t_n}(P_Tf) - \mu_{t_n}(f) &= \frac{1}{t_n}\left(\int_T^{t_n+T} P_sf(x)ds - \int_0^{t_n} P_sf(x)ds\right)\\ &= \frac{1}{t_n}\left(\int_{t_n}^{t_n+T} P_sf(x)ds - \int_0^T P_sf(x)ds\right) \leq \frac{2T\|f\|_\infty}{t_n} \xrightarrow{n\to+\infty} 0.\end{aligned}$$

On a donc pour tout $T > 0$ et $f \in \mathcal{C}_0$, $\mu_\infty(P_Tf) = \mu_\infty(f)$. Cela conclut la preuve. □

Exemple : En utilisant la fonction $V(x) = 1 + |x|^2$, on peut alors facilement montrer que si σ est bornée et si

$$\limsup_{|x|\to+\infty} \langle x, b(x)\rangle < 0,$$

alors, on a existence de mesure invariante. On peut penser par exemple au processus d'Ornstein-Uhlenbeck.

Références : Une excellente référence sur le sujet est l'ouvrage suivant :

[K12] R. Khasminskii *Stochastic stability of Differential Equations*, Second Edition, Stochastic Modelling and Applied Porbability, Springer.

Ce livre traite principalement du cas des diffusions pour les équations différentielles stochastiques mais tout ce qui y est raconté peut avoir un formalisme type semi-groupe très simple. On peut aussi renvoyer au survey de Gilles Pagès "Sur quelques algorithmes récursifs pour les probabilités numériques" (ESAIM :PS,2002) dont le théorème ci-dessus est issu et dont quelques résultats à venir se rapprochent.

5.2.3 Existence de mesures stationnaire (cadre trajectoriel)

À noter que ces fonctions de Lyapunov étaient déjà fondamentales pour prouver la non explosion du processus en temps fini. Nous introduisons l'hypothèse de récurrence :

Hypothèse 1 $(\mathbf{H_R})$: *Il existe un ensemble U ouvert borné de E de bord Γ lisse tel que*

1. $(\mathbf{H_{R1}})$ *Dans U, le processus est elliptique, par exemple la matrice de diffusion est non dégénérée de déterminant minoré.*
2. $(\mathbf{H_{R2}})$ *Pour tout compact K de $E \setminus U$, le temps moyen τ_U issu de $x \in K$ pour rentrer dans U est uniformément borné :*

$$\sup_{x \in K} \mathbb{E}_x \tau_U \leq T_K < +\infty.$$

Nous supposons dans un premier temps que les conditions $(\mathbf{H_R})$ sont satisfaites et construisons un ouvert U_1 de frontière Γ_1 vérifiant $(\mathbf{H_R})$. Dans cet ouvert U_1, on considère alors un ouvert U de bord régulier Γ tel que $U \cup \Gamma \subset U_1$. Nous renvoyons à la figure 5.1 pour une réalisation imagée.

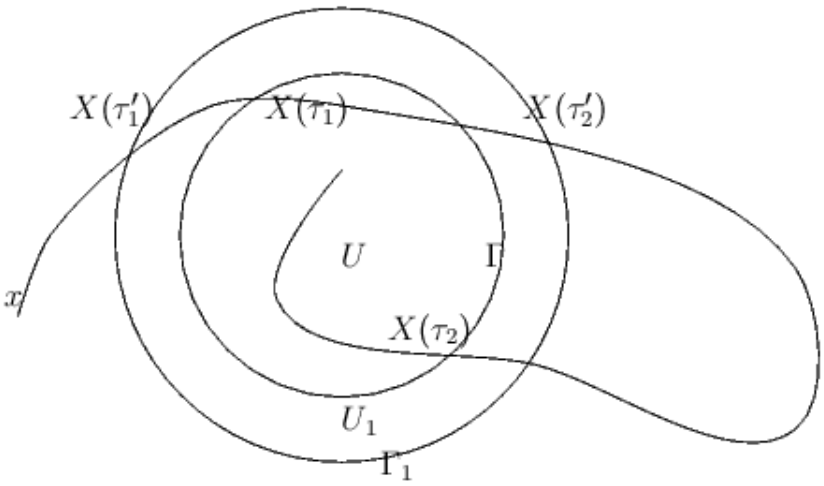

FIGURE 5.1: Construction de la chaine squelette.

On définit alors

$$\tau_0 = 0 \qquad \text{et} \qquad \tau_1' = \inf\{t \geq 0 \quad | \quad X_t \in \Gamma_1\},$$

puis par récurrence :

$$\tau_n = \inf\{t \geq \tau_n' \quad | \quad X_t \in \Gamma\} \qquad \text{et} \qquad \tau_{n+1}' = \inf\{t \geq \tau_n \quad | \quad X_t \in \Gamma_1\}.$$

Par hypothèse, on sait que le processus est récurrent irréductible sur U_1 puisque $\mathcal{L}$ y est elliptique par $(\mathbf{H_{R1}})$ et toute excursion hors de U_1 y revient irrémédiablement par $(\mathbf{H_{R2}})$. Ainsi, les temps τ_n et τ_n' sont finis presque sûrement.

Definition 5.2.3 (Chaîne squelette) *On définit la chaîne squelette par*

$$\tilde{X}_i := X(\tau_i), \quad \forall i \geq 0.$$

La semi-groupe de transition pour $\tilde{X}$ sera noté $\tilde{P}$:

$$\mathbb{E}_x f(\tilde{X}_1) = \int_{\Gamma_1} \tilde{P}(x, dy) f(y).$$

Par compacité de Γ_1, on peut démontrer un résultat proche du théorème de Doeblin.

Proposition 5.2.3 *La chaîne de Markov est irréductible et récurrente sur Γ_1. Elle admet une unique mesure invariante $\tilde{\mu}$ portée par Γ_1 et il existe $k \in (0,1)$ tel qu'uniformément en $\gamma \in \Gamma_1$:*

$$\left| \tilde{P}^n(x, \gamma) - \tilde{\mu}(\gamma) \right| \leq k^n.$$

Nous allons ensuite construire à partir de $\tilde{\mu}$ une mesure invariante pour le processus $(X_t)_{t \geq 0}$, sisi ! On définit pour tout ensemble mesurable A le temps passé dans A par le processus avant le premier retour dans Γ_1 :

$$\pi(A) := \int_{\Gamma_1} \tilde{\mu}(dx) \mathbb{E}_x \int_0^{\tau_1} 1_{X_s \in A} ds.$$

On a alors le résultat fondamental :

Theorem 5.2.2 (Existence de mesure invariante) *Sous les hypothèses* $(\mathbf{H_R})$*, le processus $(X_t)_{t \geq 0}$ possède π comme mesure invariante. De plus, cette mesure est de masse finie et peut être normalisée en probabilité invariante.*

<u>Preuve :</u> On considère f définie sur E et

$$\pi(f) = \int_{\Gamma_1} f(y) \pi(dy) = \int_{\Gamma_1} \tilde{\mu}(dx) \mathbb{E}_x \int_0^{\tau_1} f(X_u) du.$$

Fixons désormais un temps t quelconque, on a alors par la propriété de Markov que

$$\pi(P_t f) = \int_E \pi(dx) f(X_t^x) = \int_{\Gamma_1} \tilde{\mu}(dx) \mathbb{E}_x \int_0^{\tau_1} f(X_{t+s}) ds.$$

et un changement de variable $u = t + s$ montre que

$$\begin{aligned} \pi(P_t f) &= \int_{\Gamma_1} \tilde{\mu}(dx) \mathbb{E}_x \int_t^{t+\tau_1} f(X_u) du \\ &= \int_{\Gamma_1} \tilde{\mu}(dx) \mathbb{E}_x \int_0^{\tau_1} f(X_u) du + \int_{\Gamma_1} \tilde{\mu}(dx) \mathbb{E}_x \int_{\tau_1}^{\tau_1+t} f(X_u) du \\ &\quad - \int_{\Gamma_1} \tilde{\mu}(dx) \mathbb{E}_x \int_0^t f(X_u) du. \end{aligned}$$

Pour le terme du milieu entre τ_1 et $\tau_1 + t$, on peut tout conditionner par rapport à $\tilde{X}_1$ et

$$\int_{\Gamma_1} \tilde{\mu}(dx) \mathbb{E}_x \int_{\tau_1}^{\tau_1+t} f(X_u) du = \int_{\Gamma_1} \tilde{\mu}(dx) \mathbb{E}_x \underbrace{\mathbb{E}_{\tilde{X}_1} \int_0^t f(X_u) du}_{:= g(\tilde{X}_1)}. \tag{5.4}$$

Maintenant, on sait que $\tilde{\mu}$ est stationnaire pour $\tilde{X}$ donc

$$\tilde{\mu}(g) = \tilde{\mu}(\tilde{P} g),$$

c'est à dire pour toute fonction ψ mesurable et bornée :

$$\int_{\Gamma_1} \tilde{\mu}(dx)\psi(x) = \int_{\Gamma_1} \tilde{\mu}(dx)\mathbb{E}_x\psi(\tilde{X}_1). \tag{5.5}$$

En utilisant l'invariance donnée dans Equation (5.5) dans Equation (5.4) avec $\psi = g$, on obtient

$$\int_{\Gamma_1} \tilde{\mu}(dx)\mathbb{E}_x \int_{\tau_1}^{\tau_1+t} f(X_u)du = \int_{\Gamma_1} \tilde{\mu}(dx)g(x) = \int_{\Gamma_1} \tilde{\mu}(dx)\mathbb{E}_x \int_0^t f(X_u)du$$

En fin de compte, en reprenant l'expression finale obtenue pour $\pi(P_t f)$ et en supprimant ce qui doit l'être, on trouve

$$\pi(P_t f) = \int_E \mathbb{E}_x f(X_t)\pi(dx) = \int_{\Gamma_1} \tilde{\mu}(dx)\mathbb{E}_x \int_0^{\tau_1} f(X_u)du = \int_E \pi(dx)f(x) = \pi(f).$$

La mesure π est donc invariante.

Pour obtenir le caractère fini de la mesure, il suffit de calculer $\pi(E)$. Mais on sait d'après $(\mathbf{H_{R2}})$ que

$$\sup_{x\in\Gamma_1} \mathbb{E}_x\tau_1 \leq T_{U_1} < +\infty.$$

La mesure est donc finie. □

5.2.4 Controler les temps de retour dans les compacts

Où nous en sommes Nous donnons dans cette partie une méthode exploitant totalement l'outil « fonction de Lyapunov » pour obtenir que l'hypothèse $(\mathbf{H_R})$ soit satisfaite.

Le premier volet de $(\mathbf{H_{R1}})$ est une condition d'irréductibilité, nous avons vu que cela mettait en jeu des propriétés de régularité du semi-groupe (ellipticité de $\mathcal{L}$), et de controlabilité trajectorielle (on peut passer d'un point quelconque de E à un autre quelconque avec probabilité strictement positive). Aussi, nous ne reviendons pas sur $(\mathbf{H_{R1}})$. Attardons nous plutôt sur la compactification réclamée par $(\mathbf{H_{R2}})$.

La condition $\mathcal{L}V \leq \beta - \alpha V$ L'idée est de trouver une fonction V minorée (en fait positive la plupart du temps), dite fonction de Lyapunov possédant la propriété $V(x) \mapsto +\infty$ lorsque $|x| \mapsto +\infty$ telle que l'évolution de $V(X_t)$ a plutôt tendance à être bornée.

Si pour une telle fonction V pour laquelle nous avons l'existence de $\alpha > 0$ et $\beta > 0$ tels que

$$\mathcal{L}V \leq \beta - \alpha.$$

Fixons un compact K quelconque de E, comme V est coercive, on choisit et un ouvert U tel que

$$\forall x \notin U \qquad \alpha V(x) \geq 2\beta. \tag{5.6}$$

On définit τ_U le temps de rentrée dans U et $\tau_U^n = \tau_U \wedge n$ qui est un temps d'arrêt borné. La formule de Dynkin (Ito appliquée à un temps d'arrêt) montre que

$$\forall x \in K \qquad \mathbb{E}_x V(X_{\tau_U^n}) = V(x) + \mathbb{E}_x \int_0^{\tau_U^n} \mathcal{L}V(X_s)ds$$

Avant τ_U^n, le processus n'est pas dans U donc on peut appliquer Equation (5.6) pour obtenir que

$$\forall x \in K \qquad \mathbb{E}_x V(X_{\tau_U^n}) + \mathbb{E}_x \int_0^{\tau_U^n} -\mathcal{L}V(X_s)ds = V(x),$$

d'où

$$\forall x \in K \qquad V(x) \geq \beta \mathbb{E}_x \int_0^{\tau_U^n} 1 ds = \beta \mathbb{E}_x \tau_U^n.$$

Nous obtenons alors par convergence monotone la borne :

$$\sup_{x \in K} \mathbb{E}_x \tau_U \leq \frac{\sup_{x \in K} V(x)}{\beta}. \tag{5.7}$$

Nous avons donc établi le théorème

Theorem 5.2.3 (Existence de mesure invariante) *S'il existe V fonction de Lyapunov (vérifiant $\min V > 0$, coercive) telle que $\mathcal{L}V \leq \beta - \alpha V$ pour $(\alpha, \beta) \in \mathbb{R}_+^2$, alors il existe U un ouvert tel que pour tout compact K :*

$$\sup_{x \in K} \mathbb{E}_x \tau_U \leq C_K.$$

En particulier, si $\mathcal{L}$ est elliptique sur U il existe une mesure invariante pour $(X_t)_{t \geq 0}$.

5.3 Unicité de la probabilité invariante

Ce résultat d'unicité est à rechercher dans le contexte de l'irréductibilité de la transition définie dans $(P_t)_{t \geq 0}$. En effet, pour les chaînes de Markov, c'est déjà l'irréductibilité de la transition Q qui permet de donner des résultats d'unicité des distributions stationnaires (voir le premier théorème du cours).

Nous allons volontairement nous placer dans un cadre simplifié où le générateur infinitésimal $\mathcal{L}$ est elliptique et $E = \mathbb{R}^d$ et nous admettons que le semi-groupe P_t agit sur la mesure initiale μ en lui conférant une densité par rapport à la mesure de Lebesgue sur E. Cette densité au temps t est notée $p_t(x, y)$. Ce genre de résultat un petit peu anecdotique par rapport à ce cours est par exemple vrai dès que le système dynamique est elliptique (covariance inversible devant le mouvement brownien dans une e.d.s.) ou hypo-elliptique (conditions de Hormandër satisfaites localement).

La formule d'Ito s'écrit pour toute fonction $f \in \mathcal{D}(\mathcal{L})$:

$$\mathbb{E}_x [f(X_t)] = \int_0^t \mathbb{E}_x \mathcal{L} f(X_s) ds.$$

Par ailleurs, on a

$$\mathbb{E}_x [f(X_t)] = \int_E p_t(x, y) f(y) dy \quad \text{et} \quad \mathbb{E}_x \mathcal{L} f(X_s) = \int_E \mathcal{L} f(y) p_t(x, y) dy.$$

En différentiant par rapport au temps et en notant $\mathcal{L}^*$ l'adjoint de $\mathcal{L}$ par rapport à la mesure de Lebesgue, on obtient que la mesure au temps t vérifie l'e.d.p. :

$$\partial_t p_t(x, y) = \mathcal{L}^* p_t(x, y).$$

Cette équation est l'équation backward de Kolmogorov (ou de Fokker-Planck). $\mathcal{L}^*$ étant elliptique, on en déduit alors que $p_t(., .)$ est une fonction $\mathcal{C}^\infty$ en tout temps $t > 0$. Par ailleurs, étant donnée une mesure invariante π (on confond ici mesure et densité par rapport à la mesure de Lebesgue) pour $\mathcal{L}$, elle est également strictement positive p.s. puisqu'on peut vérifier la relation

$$\pi(z) = \int_E p_t(x, z) \pi(x) dx.$$

Nous venons d'établir la proposition :

Proposition 5.3.1 *Si $\mathcal{L}$ est elliptique ou hypo-elliptique, et si μ est la loi initiale de X_0, alors la mesure $P_t\mu$ est à densité par rapport à la mesure de Lebesgue, de densité p_t qui est $\mathcal{C}^\infty(E)$. Par ailleurs, si on initialise en $X_0 = x$, la densité $(x, y) \mapsto p_t(x, y)$ est $\mathcal{C}^\infty(E \times E)$. Cette densité vérifie l'équation backward de Kolmogorov :*

$$\partial p_t = \mathcal{L}^* p_t.$$

Par ailleurs, si on a une dynamique $dx_t = b(x_t)dt + dB_t$, alors on a

$$\forall t > 0 \quad \forall (x, y) \in E^2 \qquad p_t(x, y) > 0.$$

Nous établissons maintenant un résultat d'unicité de mesure invariante lorsque le semi-groupe est irréductible.

Lemma 5.3.1 *Si une mesure π vérifie $P_t\pi \leq \pi$ pour tout $t > 0$, alors celle-ci est invariante.*

Preuve : La démonstration est très simple : on prend un ensemble mesurable A et on remarque que $(P_t\pi)(A) \leq \pi(A)$ et $(P_t\pi)(A^c) \leq \pi(A^c)$. Cela implique bien entendu l'égalité des probabilités et donc que π est invariante. □

On passe maintenant au résultat d'unicité de mesure invariante qui est vrai dans des situations un peu plus générale que ce qui est énoncé ici.

Theorem 5.3.1 (Théorème d'unicité de Doob) *Si le semi-groupe est elliptique et irréductible, alors il existe au plus une unique mesure invariante.*

Preuve : On considère deux distributions invariantes π_1 et π_2 pour le semi-groupe. On sait que ces deux distributions sont absolument continues par rapport à la mesure de Lebesgue. On identifie ces deux mesures avec leurs densités (qui sont strictement positives sur E) et on écrit

$$\pi_1 = h\pi_2.$$

On suppose qu'il existe un x de E tel que $h(x) = a > 1$. On considère $\nu_a = (h \wedge \underbrace{(1 + a)/2}_{:=\tilde{a}})\pi_2$ et cette distribution satisfait

$$P_t\nu_a = P_t(h \wedge a)\pi_2 \leq P_t h\pi_2 = P_t\pi_1 = \pi_1 = h\pi_2.$$

De même,

$$P_t\nu_a \leq P_t\tilde{a}\pi_2 = \tilde{a}P_t\pi_2 = \tilde{a}\pi_2.$$

Du coup, on a bien que

$$\forall t > 0 \qquad P_t\nu_a \leq \nu_a,$$

et ν_a est une mesure invariante d'après la proposition précédente. Ainsi, le semi-groupe laisse invariante $\mu_a = (1 + a)/2\pi_1 - \nu_a$ mais $\mu_a = [\frac{1+a}{2} - h]_+\pi_1$. Cette mesure peut être renormalisée en mesure de probabilité mais elle contient également un point x pour lequel $\mu_a(x) = 0$. C'est impossible puisque $P_t\mu_a = \mu_a$ qui implique la stricte positivité partout. On obtient alors que h est identiquement égale à 1. Cela conclut la preuve de l'unicité. □

Un cadre d'application important est le suivant :

Corollary 5.3.1 *Soit (P_t) un semi-groupe sur $\mathbb{R}^d$. Supposons qu'il existe $t_0 > 0$ tel que $P_{t_0}(x, dy) = p_{t_0}(x, y)\lambda_d(dy)$. Alors, on a unicité de la mesure invariante μ dès que dy-pp, pour tout $x \in \mathbb{R}^d$, $p_t(x, y) > 0$. De plus, μ est équivalente à la mesure de Lebesgue.*

Ce résultat n'est pas exactement un corollaire car les hypothèses ci-dessus assurent l'irréductibilité mais pas la propriété de Feller forte. Néanmoins, les outils de la preuve sont proches. On pourra vérifier en exercice que sous l'hypothèse précédente, la preuve du théorème fonctionne encore.

Exemple important : Supposons que $(X_t)_{t\geq 0}$ soit une diffusion uniformément elliptique : $dX_t = b(X_t)dt + \sigma(X_t)dB_t$ avec $b : \mathbb{R}^d \to \mathbb{R}^d$ et $\sigma : \mathbb{R}^d \to \mathbb{M}_{d,d}$ régulières et $\sigma\sigma^* \leq \lambda_0 I_d$ (minoration au sens des matrices symétriques), où $\lambda_0 > 0$. Alors,(X_t) admet une densité strictement positive régulière en (x, y). C'est évidemment le cas si σ est une matrice constante et non dégénérée.

5.4 Calcul explicite de mesures invariantes, exemples

5.4.1 Calcul explicite

On part de l'hypothèse que le semi-groupe (P_t) admet une unique mesure invariante μ de densité p (suffisamment) régulière par rapport à la mesure de Lebesgue. En notant $\langle f, g\rangle = \int f(x)g(x)\lambda_d(dx)$, *i.e.* le produit scalaire sur $L^2(\mathbb{R}^d, \lambda_d)$, on a :

$$\int \mathcal{L}f d\mu = \int \mathcal{L}f(x)p(x)\lambda_d(dx) = \langle \mathcal{L}f, p\rangle = \langle f, \mathcal{L}^*p\rangle$$

où $\mathcal{L}^*$ est l'adjoint de $\mathcal{L}$. D'après la proposition 5.2.1, $\langle f, \mathcal{L}^*p\rangle = 0$ si $f \in \mathcal{D}(\mathcal{L})$ et satisfait $(\mathbf{H_f})$. Si la classe de fonctions est suffisamment grande, alors on en déduit que p est l'unique solution de

$$\mathcal{L}^*p = 0.$$

Pour une diffusion en dimension 1 ergodique, cette équation admet une équation explicite lorsque $\sigma(x) \neq 0$ et b et σ sont $\mathcal{C}^1$.

Exercice : Calculer la densité de la mesure invariante pour une diffusion unidimensionnelle ergodique (sous les hypothèses précédentes). On commencera par montrer (par intégrations par partie) que pour toute fonction $\mathcal{C}^2$ à support compact, $\langle \mathcal{L}f, p\rangle = \langle f, \mathcal{L}^*p\rangle$ puis on résoudra l'équation $\mathcal{L}^*p = 0$.

Example 5.4.1 (Semi-groupe d'Ornstein-Uhlenbeck) *Il s'agit du système diffusif ergodique le plus simple :*

$$dX_t = -\lambda X_t dt + \sigma dB_t, \qquad \lambda, \sigma > 0.$$

On suppose que (X_t) est à valeurs dans $\mathbb{R}^d$. C'est un système avec drift linéaire vers l'origine, d'intensité λ. On vérifiera la propriété de non explosion avec une fonction judicieusement choisie (un polynôme en x par exemple). Par ailleurs, on identifiera la mesure invariante à l'aide du critère de la section précédente au travers du calcul de l'adjoint du générateur du semi-groupe d'Ornstein-Uhlenbeck.

Enfin, il est évidemment possible d'écrire que X_t est en réalité égal à :

$$X_t = X_0 e^{-\lambda t} + \int_0^t e^{\lambda(s-t)} dB_s,$$

Montrer que $\pi = \mathcal{N}(0, \sigma^2/(2\lambda))$ est l'unique mesure invariante du système (On pourra soit regarder la convergence des moments, soit montrer que cette mesure est solution de l'équation $\int \mathcal{L}f d\pi = 0$ pour tout f $\mathcal{C}^2$ à support compact).

Example 5.4.2 (Diffusions gradient) *Qu'en est-il pour l'EDS suivante*

$$dX_t = -\nabla U(X_t)dt + \sigma dB_t$$

avec $\sigma > 0$, $U : \mathbb{R}^d \to \mathbb{R}$ $\mathcal{C}^1$ *telle que* $\lim_{|x|\to+\infty} U(x) = +\infty$. *Montrer que si* $\lim_{|x|\to+\infty} |\nabla U(x)|^2 = +\infty$ *lorsque* $|x| \to +\infty$, *alors* $p(x) = Z^{-1}\exp(-U(x)/(2\sigma^2)$ *(où* $Z = \int_{\mathbb{R}^d} \exp(-U(x)/(2\sigma^2)dx)$ *est l'unique mesure invariante.*

5.5 Vitesse de convergence à l'équilibre

Dans cette section, on suppose que l'on a existence et unicité de la mesure invariante π et on aborde le vaste sujet de la convergence vers la mesure invariante. Par le théorème de Birkhoff énoncé précédemmment, on a un premier résultat de convergence en moyenne trajectorielle (rappelons cependant que ce résultat n'est vrai que $\pi(dx)$ presque partout). On va ici proposer quelques résultats de convergence en loi.

5.5.1 Forme de Dirichlet

Une documentation archi-complète sur ce très vaste domaine se trouve dans

[ABC+] C.Ané, S.Blachère, D.Chafaï, P.Fougères, I.Gentil, F.Malrieu, C.Roberto, G.Scheffer *Sur les inégalités de Sobolev logarithmiques*, Société Mathématique de France, 2000

Vitesse exponentielle à l'équilibre

Definition 5.5.1 (Forme de Dirichlet) *Étant donné un opérateur* $\mathcal{L}$ *qui laisse invariante* μ, *on définit*

$$\forall f \in \mathcal{D}(\mathcal{L}) \qquad \Gamma(f,g) = \frac{1}{2}[\mathcal{L}(fg) - f\mathcal{L}(g) - g\mathcal{L}(f)]$$

On définit également $\Gamma(f) = \Gamma(f,f)$.

Proposition 5.5.1 *On a les propriétés suivantes :*
- $\forall f \in \mathcal{D}(\mathcal{L}) \qquad \Gamma(f) \geq 0$
- $\int \Gamma(f)d\mu = -\int f\mathcal{L}(f)d\mu$

Proof : On démontre la première propriété en remarquant que P_t est un opérateur positif : toute fonction positive est transformée en fonction positive sous P_t. Ainsi :

$$P_t((f-a)^2) \geq 0.$$

Par ailleurs, on a

$$P_t((f-a)^2) = P_t(f^2) - 2aP_t(f) + a^2$$

Aussi, le discriminant du polynôme en a précédent est négatif. Cela signifie donc que

$$P_t(f^2) \geq P_t(f)^2.$$

Puis, on peut remarquer que

$$\lim_{t\to 0} \frac{P_t(f^2) - f^2}{t} = \mathcal{L}(f^2)$$

alors que

$$\lim_{t\to 0} \frac{P_t(f)^2 - f^2}{t} = \lim_{t\to 0} \frac{P_t(f) - f}{t}(P_t(f) + f) = 2f\mathcal{L}(f)$$

puisque $P_t(f) \longrightarrow f$ lorsque $t \longrightarrow 0$. Aussi, on obtient que

$$\lim_{t \longrightarrow 0} \frac{P_t(f^2) - P_t(f)^2}{2t} = \Gamma(f)$$

et donc $\Gamma(f) \geq 0$.
La seconde propriété résulte du fait que $\int \mathcal{L}(f) d\mu = 0$ pour toute fonction f lorsque $\mathcal{L}^\star(\mu) = 0$.

Definition 5.5.2 (Energie) *La fonction d'énergie associée à la forme de Dirichlet est définie par*

$$\mathcal{E}(f) = -\int f \mathcal{L}(f) d\mu = \int \Gamma(f) d\mu.$$

Dans ce contexte, on s'intéressera à une inégalité fonctionnelle de type Poincaré. Cette inégalité mettra en jeu la forme de Dirichlet définie précédemment et un terme de variance quantifiant la vitesse de convergence à l'équilibre. Nous allons spécifier ce lien dans le contexte des diffusions définies par

$$dX_t = b(X_t)dt + dB_t.$$

Dans ce cas, le générateur infinitésimal est donné par

$$\mathcal{L}f = \langle b(x), \nabla f(x) \rangle + \frac{1}{2}\Delta f.$$

De ce fait, l'opérateur $\Gamma(f)$ est donné par

$$\frac{1}{2}[\mathcal{L}(f^2) - 2f\mathcal{L}(f)] = \frac{1}{2}\left[2f\langle b, \nabla f\rangle + \frac{1}{2}\Delta(f^2) - 2f\langle b, \nabla f\rangle - f\Delta f\right] = \frac{< \nabla f, \nabla f >}{2},$$

le dernier terme étant obtenu en calculant explicitement $\Delta(f^2)$.

5.5.2 Diffusion de Kolmogorov

La diffusion générale la plus célèbre est assurément celle définie au travers de l'équation différentielle stochastique

$$dX_t = -\nabla U(X_t)dt + dB_t, \tag{5.8}$$

où U désigne un *potentiel* coercif, minoré (par 0). Le générateur infinitésimal de ce processus est

$$\mathcal{L}(f) = -\langle \nabla f, \nabla U \rangle + \frac{1}{2}\Delta f$$

Difficile de donner un ouvrage de synthèse sur cette équation. On pourra consulter avec profit les ouvrages.

[ABC+] C.Ané, S.Blachère, D.Chafaï, P.Fougères, I.Gentil, F.Malrieu, C.Roberto, G.Scheffer *Sur les inégalités de Sobolev logarithmiques*, Société Mathématique de France, 2000.

[B94] D.Bakry, *L'hypercontractivité et son utilisation en théorie des semigroupes*, Lectures on Probability theory, Lecture Notes in Math. 1581, 1994

[R99] G. Royer, *Une initiation aux iné ?galités de Sobolev logarithmiques*, S.M.F., Paris, 1999.

Cette diffusion est intimement à la dynamique décrivant le recuit simulé. Nous commençons à donner des éléments importants liés à cette diffusion fondamentale. On effectue l'hypothèse que U est de classe $\mathcal{C}^2$ et tel que

$$\lim_{|x| \longrightarrow +\infty} U(x) = +\infty \qquad \text{et} \qquad |\nabla U|^2 - \Delta U \quad \text{est minoré.}$$

Par ailleurs, nous définissons la condition de rappel à l'infini :

$$\exists a \in \mathbb{R}_+ \exists b \in \mathbb{R} \quad \forall x \in \mathbb{R}^p \qquad \langle x, \nabla U(x) \rangle \geq a|x|^2 - b$$

Nous obtiendrons l'existence et l'unicité forte des solutions de ce processus ainsi que la non explosion en temps fini en utilisant par exemple la fonction de Lyapunov $V(x) = |x|_2^2$ puisque

$$\mathcal{L}(V) = -2 < \nabla U(x), x \rangle + |x|^2 \leq 2b - 2aV.$$

En appliquant alors le théorème 5.2.1, ce dernier point nous assure également que le processus admet au moins une distribution invariante. Par ailleurs, le semi-groupe étant elliptique, nous avons également unicité de la mesure invariante associée à Equation (5.8). Nous obtenons alors le théorème suivant.

Theorem 5.5.1 *Le processus $(X_t)_{t\geq 0}$ définit par Equation (5.8) possède une unique distribution invariante π donnée par*

$$\pi(x) \propto e^{-2U(x)}.$$

Proof : La seule chose qu'il reste à prouver concerne l'invariance de π, définie à une constante multiplicative près. Nous commençons par démontrer que π est réversible pour $\mathcal{L}$, c'est-à-dire que $\mathcal{L}$ est auto-adjoint dans $\mathbb{L}^2(\pi)$.

$$\forall (f,g) \in \mathbb{L}^2(\pi) \qquad < f, \mathcal{L}(g) >_\pi = \int f \mathcal{L}(g) d\pi = \int f \Delta g e^{-U/2} + \int f < \nabla U, \nabla g > e^{-2U}.$$

On intègre alors par parties le premier terme du second membre en remarquant que $\nabla e^{-2U} = -2e^{-2U} \nabla U$. On obtient alors que

$$< f, \mathcal{L}(g) >_\pi = -\int \nabla f \nabla g d\pi =< \mathcal{L}(f), g >_\pi .$$

Cette relation de symétrie implique alors immédiatement que π est invariante :

$$\forall f \in \mathbb{L}^2(\pi) \qquad \int \mathcal{L}(f) d\pi =< 1, \mathcal{L}(f) >_\pi =< \mathcal{L}(1), f >_\pi = 0.$$

On obtient ainsi la preuve du théorème. □

5.5.3 Inégalité de Poincaré et opérateurs auto-adjoints

Nous commençons par établir un résultat classique lorsqu'on sait démontrer une inégalité de Poincaré sur la mesure stationaire.

Definition 5.5.3 (Inégalité de Poincaré) *Soit π la mesure **stationnaire** associée au semi-groupe de générateur $\mathcal{L}$. La mesure satisfait une inégalité de trou spectral (ou de Poincaré) s'il existe une constante C_P telle que*

$$\forall f \in \mathcal{D}(\mathcal{L}) \qquad Var_\pi(f) = \int_E [f - \pi(f)]^2 d\pi \leq C_P \mathcal{E}(f,f).$$

Dans le contexte de la diffusion de Kolmogorov Equation (5.8), cette inégalité s'écrit :

$$\forall f \in \mathbb{L}^2(\pi) \qquad \int_{\mathbb{R}^p} [f(x) - \pi(f)]^2 d\pi(x) \leq C_P \int_{\mathbb{R}^p} |\nabla f(x)|^2 d\pi(x).$$

Dans le cas précis où l'opérateur $\mathcal{L}$ est auto-adjoint dans $\mathbb{L}^2(\pi)$, on peut alors diagonaliser $\mathcal{L}$. Tout d'abord, il est immédiat de voir que 0 est valeur propre associée à la fonction constante 1.

Notons par ailleurs λ une valeur propre de $-\mathcal{L}$ non nulle, de vecteur propre associé f tel que $\pi(f) = 0$. Alors :

$$\int_E f^2 d\pi = \frac{1}{\lambda} \int_E -\mathcal{L}(f) f d\pi.$$

Par ailleurs, nous avons vu que $2\Gamma(f,f) = \mathcal{L}(f^2) - 2f\mathcal{L}(f)$ et que

$$\begin{aligned} \int_E -\mathcal{L}(f) f d\pi &= \int_E \Gamma(f,f) d\pi - \frac{1}{2} \int_E \mathcal{L}(f^2) d\pi \\ &= \mathcal{E}(f,f) - \langle \mathcal{L}(f^2), 1 \rangle_\pi \\ &= \mathcal{E}(f,f) - \langle f^2, \mathcal{L}(1) \rangle_\pi \\ &= \mathcal{E}(f,f), \end{aligned}$$

car $\mathcal{L}(1) = 0$.

Remark 5.5.1 *On peut aussi voir cette dernière annulation en utilisant l'adjoint de $\mathcal{L}$ pour la mesure de Lebesgue pour lequel $\mathcal{L}^*(\pi) = 0$ par stationnarité.*

Nous venons donc d'établir que si $\mathcal{L}(f) = -\lambda f$, alors

$$Var_\pi(f) = \frac{1}{\lambda} \mathcal{E}(f,f).$$

Par conséquent, si $\pi, \mathcal{L}$ vérifient une inégalité de Poincaré de constante C_P, on a alors

$$\frac{1}{\lambda} \leq C_P.$$

Cela signifie donc que les valeurs propres de $-\mathcal{L}$ satisfont la propriété de « trou spectral » :

$$Sp(-\mathcal{L}) \subset \{0\} \cup \left[\frac{1}{C_p}, +\infty \right[.$$

5.5.4 Convergence exponentielle

Nous pouvons désormais démontrer le théorème de convergence exponentielle du semi-groupe vers π. L'objet que nous venons naturellement d'introduire est la variance d'une fonction f. Afin de démontrer une convergence en loi du processus (X_t) vers π, nous allons quantifier la proximité de la loi de (X_t) à π. Un critère populaire est donné par la distance $\mathbb{L}^2$.

$$\phi(t) := \int_{\mathbb{R}^p} [E_x[f(X_t)] - \pi(f)]^2 d\pi(x).$$

Elle mesure comment la loi de X_t partant de x s'approche de π, et ce en intégrant sur l'espace pondéré par π. Il s'agit donc de calculer

$$\phi(t) = Var_\pi(P_t f),$$

car $P_t f(x) = \mathbb{E}_x f(X_t)$. Nous démontrons alors le théorème suivant.

Theorem 5.5.2 (Ergodicité par trou spectral, cas symétrique) *Soit P_t un semi-groupe de générateur $\mathcal{L}$ et mesure invariante π, tels que $\mathcal{L}$ est symétrique dans $L^2(\pi)$. Il y a* **équivalence** *entre les deux propositions suivantes.*

i) $\mathcal{L}$ satisfait une inégalité de Poincaré de constante C_P.

ii) Il y a convergence exponentielle :

$$\forall f \in \mathcal{D}(\mathcal{L}) \qquad Var_\pi(P_t f) \leq e^{-2t/C_p} Var_\pi(f).$$

Preuve : On démontre tout d'abord que $i) \Rightarrow ii)$. On suppose que $\pi(f) = 0$ sans perte de généralité. On définit

$$\phi(t) := \int_E [P_t f]^2 d\pi = Var_\pi(P_t f).$$

On a alors

$$\begin{aligned} \phi'(t) &= \int_E 2\mathcal{L}(P_t f)(P_t f) d\pi \\ &= -2\mathcal{E}(P_t f, P_t f). \end{aligned}$$

L'inégalité de Poincaré implique alors que

$$\phi'(t) \leq -\frac{2}{C_P}\phi(t),$$

et on conclut en appliquant le lemme de Gronwall :

$$Var_\pi(P_t f) \leq e^{-\frac{2}{C_P}t} Var_\pi(f).$$

Réciproquement, montrons que $ii) \Rightarrow i)$ et supposons que l'on ait l'ergodicité exponentielle. Pour $t = 0$, il est facile de voir que les deux côtés de l'inégalité sont égaux. Ainsi, la dérivée en $t = 0$ du terme de droite est supérieure à la dérivée en $t = 0$ du terme de gauche :

$$\phi'(0) \leq -\frac{2}{C_P} Var_\pi(f).$$

Mais un calcul rapide montre que

$$\phi'(0) = -2\mathcal{E}(f, f).$$

Cela aboutit directement à l'inégalité de Poincaré :

$$Var_\pi(f) \leq C_P \mathcal{E}(f, f)$$

Le théorème est donc démontré. □

Tensorisation Enfin, il y a une façon de déduire des inégalités de Poincaré si on arrive à décomposer le générateur $\mathcal{L}$. Cela peut se faire par le biais de la propriété de « tensorisation ». Considérons $\mathcal{L}_1$ et $\mathcal{L}_2$ deux générateurs infinitésimaux de semi-groupe $(P_t^1)_{t\geq 0}$ sur (E_1, π_1) et $(P_t^2)_{t\geq 0}$ sur (E_2, π_2). On étudie le semi-groupe sur l'espace produit $E_1 \times E_2$ pour la mesure $\pi = \pi_1 \otimes \pi_2$ donné par

$$P_t((x_1, x_2), (dx_1, dx_2)) = P_t^1(x_1, dx_1) P_t^2(x_2, dx_2).$$

On montre facilement (exercice...) que le générateur est alors

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2.$$

Cette notation doit être comprise ainsi : pour une fonction f de deux variables (x_1, x_2) on calcule $\mathcal{L}$ en appliquant $\mathcal{L}_1$ (resp. $\mathcal{L}_2$) à x_2 (resp. x_1) fixée et on additionne les résultats. L'opérateur carré du champ Γ est donc donné par $\Gamma = \Gamma_1 + \Gamma_2$. On a alors la tensorisation de la variance :

Proposition 5.5.2

$$\forall f \in \mathcal{D}(\mathcal{L}) \qquad Var_\pi(f) \leq \int_E [Var_{\pi_1}(f) + Var_{\pi_2}(f)] d\pi$$

On étudie la différence des deux termes :

$$\begin{aligned}\int_E [Var_{\pi_1}(f) + Var_{\pi_2}(f)] - Var_\pi(f) &= \int_E \left(\int_{E_1} [f(.,x_2) - \pi_1(f(.,x_2)]^2 d\pi_1 \right) d\pi \\ &+ \int_E \left(\int_{E_2} [f(x_1,.) - \pi_2(f(x_1,.)]^2 d\pi_2 \right) d\pi \\ &- \int_E [f(.,.) - \pi(f)]^2 d\pi.\end{aligned}$$

Le théorème de Tonnelli permet alors d'écrire que

$$\int_E [Var_{\pi_1}(f) + Var_{\pi_2}(f)] - Var_\pi(f) = \int_E f^2 d\pi - \int \left[\int f d\pi_1 \right]^2 d\pi - \int \left[\int f d\pi_2 \right]^2 d\pi + \left[\int f d\pi \right]^2.$$

On vérifiera alors que le dernier terme se re-écrit en fait

$$\int (f - \pi_1(f) - \pi_2(f) + \pi(f))^2 \, d\pi.$$

Ce dernier terme est bien sûr toujours positif, ce qui démontre la proposition. □

On déduit immédiatement de cette tensorisation l'inégalité de Poincaré pour le semi-groupe $(P_t)_{t \geq 0}$ décrivant la dynamique « produit ».

Theorem 5.5.3 (Tensorisation de l'inégalité de Poincaré) *Supposons que les mesures π_1 et π_2 vérifient une inégalité de Poincaré de constante C_1 (resp. C_2) pour l'opérateur carré du champ Γ_1 (resp. Γ_2) associé au générateur $\mathcal{L}_1$ (resp. $\mathcal{L}_2$). Alors il en est de même pour $\pi = \pi_1 \otimes \pi_2$ et $\Gamma = \Gamma_1 + \Gamma_2$, avec la constante $C_P = max(C_1, C_2)$:*

$$\forall f \in \mathcal{D}(\mathcal{L}) \qquad Var_\pi(f) \leq max(C_1, C_2) \int \Gamma(f,f) d\pi.$$

Défaut (?) Bien entendu, la structure spectrale est fondamentale dans toute cette partie, en particulier le caractère symétrique de $\mathcal{L}$ dans $L^2(\pi)$ est incontournable. Même si ce cas symétrique se rencontre fréquemment, le monde est parfois mal fait et des exemples très intéressants mettent en jeu des dynamiques où $\mathcal{L}$ n'est pas auto-adjoint. Dans ce cas, il est possible d'attaquer les problèmes de stabilisation du système avec d'autres outils, plus trajectoriels. Une façon de procéder est décrite dans le paragraphe suivant.

Par ailleurs, les résultats fournis par une analyse spectrale sont parfois assez difficiles à interpréter d'un point de vue trajectoriel. Ils sont en revanche remarquablement précis.

Chapitre 6

Introductions aux méthodes bayésiennes

Nous présentons dans ce chapitre quelques éléments probabilistes ayant des applications en statistiques bayésiennes.

6.1 Paradigme Bayésien

6.1.1 Modèle Statistique

Definition 6.1.1 (Modèle) *On se place dans un espace probabilisé paramétrique classique : on considère un espace probabilisé* $(\mathcal{X}, \mathcal{B}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ *et les données observées sont* X*.*

Le but de l'analyse statistique est de faire de l'inférence sur , c'est-à-dire décrire un phénomène passé ou à venir dans un cadre probabiliste. L'idée centrale de l'analyse bayésienne est de considérer le paramètre inconnu θ comme aléatoire : l'espace des paramètres Θ est muni d'une probabilité π tel que (Θ, A, π) est un espace probabilisé. Nous noterons $\theta \sim \pi$. π est appelée loi ***a priori***. Intuitivement et en termes informationnels, elle détermine ce qu'on sait et ce qu'on ne sait pas avant d'observer X.

Definition 6.1.2 (Modèle dominé) *Le modèle est dit dominé s'il existe une mesure commune dominante* μ*, c'est- à-dire pour tout* θ*,* $\mathbb{P}_\theta$ *admet une densité par rapport à* μ *:*

$$p(X|\theta) = p_\theta(X) = \frac{d\mathbb{P}_\theta}{d\mu}(X).$$

Cette fonction $p(X|\theta) = p_\theta(X)$, vue comme une fonction de θ une fois qu'on a observé un tirage de X, est appelée vraisemblance du modèle. C'est la loi de X conditionnellement à θ.

6.1.2 Loi *a posteriori*

Definition 6.1.3 (Loi *a posteriori*) *Dans le cas d'un modèle dominé, la loi jointe de* (X, θ) *s'écrit*

$$\lambda_\pi(X, \theta) = p(X|\theta)d\pi(\theta) = p(X|\theta)\pi(\theta)d\nu(\theta),$$

la dernière égalité étant valable dans le cas où l'a priori π *est absolument continu par rapport à* ν*, la mesure de Lebesgue sur* Θ*. La loi a posteriori est définie par sa densité :*

$$d\pi(\theta|X) = \frac{f(X|\theta)\pi(\theta)}{\int_\Theta f(X|\alpha)\pi(\alpha)d\nu(\alpha)} \tag{6.1}$$

La quantité m_π définie par

$$m_\pi(X) := \int_\Theta f(X|\theta)d\pi(\theta),$$

est la loi marginale de X et est une constante de normalisation de la loi a posteriori, indépendante de θ.

Nous travaillerons donc très régulièrement à une constante multiplicative près :

$$\pi(\theta|X) \propto p(X|\theta)\pi(\theta).$$

Nous ajoutons que par construction la loi *a posteriori* est absolument continue par rapport à la loi *a priori* et bien sûr, la loi *a posteriori* est une mesure aléatoire sur Θ dépendant des observations X relevées.

Example 6.1.1 (Loi gaussienne) *Voici un petit exemple concernant la loi gaussienne : on se donne une famille $(\mathbb{P}_\theta)_{\theta\in\Theta}$ où θ indexe la moyenne d'une loi gaussienne de variance connue σ^2. Ainsi, $\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2)$, et on munit $\Theta = \mathbb{R}$ de la loi a priori gaussienne*

$$\pi(\theta) = \frac{e^{-\frac{\theta^2}{2\tau^2}}}{\sqrt{2\pi}\tau}.$$

$$\pi_n(\theta|X) \propto \left(\prod_{j=1}^n e^{-\frac{(X_j-\theta)^2}{2\sigma^2}}\right) \times e^{-\frac{\theta^2}{2\tau^2}}$$

Un rapide calcul montre alors qu'en réalité :

$$\pi_n(\theta|X) \propto e^{-n\frac{(\bar{X}_n-\theta)^2}{2\sigma^2}} \times e^{-\frac{\theta^2}{2\tau^2}},$$

où $\bar{X}_n$ est la moyenne empirique des échantillons. Un dernier calcul aboutit à

$$\pi_n(\theta|X) = \mathcal{N}\left(\bar{X}_n\frac{\tau^2}{\tau^2+\sigma^2/n}, \frac{\sigma^2/n}{\sigma^2/n+\tau^2}\right).$$

Remark 6.1.1 (Consistance de l'*a posteriori*) *Cet exemple est instructif puisqu'il donne une information importante : lorsque n grandit et tend vers $+\infty$, la mesure a posteriori se concentre autour de $\bar{X}_n$ avec une variance en σ^2/n, et ce, quel que soit τ, σ ou θ_0. Ainsi, la mesure a posteriori sur θ qui est dérivée d'un a priori arbitraire sur la moyenne de X se concentre vers le bon θ_0 à vitesse $\sqrt{n}$ au sens où la masse de l'a posteriori à l'extérieur d'une boule $B(\theta_0, \sqrt{n})$ tend vers 0 presque sûrement :*

$$\pi_n\left(B(\theta_0, \sqrt{n})^c|X_1, \ldots, X_n\right) \longrightarrow 0 \qquad \textit{lorsque} \qquad n \longrightarrow +\infty.$$

6.2 Consistance bayésienne

Cette dernière remarque est fondamentale pour la suite de ce chapitre et justifie (à mon sens) l'utilisation de méthodes d'estimations bayésiennes puisque celles-ci sont validées au sens fréquentiste lorsque le nombre d'observations grandit.

6.2.1 Formulation du résultat

Les références historiques importantes sur ces questions de consistance bayésienne sont assez anciennes pour les situations paramétriques.

[IK79] I. A. Ibragimov, R. Z. Khasminskii, *Statistical estimation : asymptotic theory*; translated by Samuel Kotz, New York : Springer-Verlag, 1979.

[lC86] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*, Springer, 1986.

[S65] L. Schwartz, *On Bayes procedure.* Z. Wahrscheinlichkeitstheorie, No. 4, pp. 10-26, 1965.

D'autres sont plus récentes dans les situations où Θ n'est plus un espace paramétrique mais de dimension « infinie ».

[GGvdW00] S. Ghosal, J. K. Ghosh, A.W. van der Vaart, *Convergence rates of posterior distributions*, The Annals of Statistics, Volume 28, Number 2, 2000.

[WLP07] S. G. Walker, A. Lijoi and I. Prunster, *On rates of convergence for posterior distributions in infinite dimensional models*, The Annals of Statistics, Vol. 35, No. 2, 738-746, 2007.

Les résultats de consistance bayésienne s'expriment en général tous sous la forme suivante : on définit une boule autour de θ_0 de rayon ϵ_n au sens d'une métrique sur les lois de probabilités :

$$B_d(\theta_0, \epsilon_n) := \{\theta \in \Theta \quad | \quad d(\mathbb{P}_\theta, \mathbb{P}_{\theta_0}) \leq \epsilon_n\}.$$

Pour cette boule, on obtient alors en général des résultats sous la forme :

$$\pi_n\left(B_d(\theta_0, \epsilon_n)\right) \longrightarrow 1 \quad p.s. \qquad \text{lorsque} \qquad n \longrightarrow +\infty.$$

Ainsi, les résultats concernent avant tout des lois de probabilités $\mathbb{P}_\theta$ et non les paramètres eux-mêmes. Par ailleurs, la distance joue un rôle fondamental et la plupart du temps, c'est la distance au sens de Kullback qui est utilisée. Enfin, la vitesse de contraction est quantifiée par la vitesse de décroissance de $(\epsilon_n)_{n\in\mathbb{N}}$ lorsque n tend vers $+\infty$.

6.2.2 Cas où Θ est fini

Comme toujours, c'est souvent le cas en apparence le plus simple qui est le plus informatif. Revenons tout d'abord à la nature de la loi *a posteriori*. En réalité, il est facile de voir que cette loi *a posteriori* suit une Markovienne. En effet, notons $(X_1, \ldots, X_n, \ldots)$ une suite de v.a. i.i.d. de loi $\mathbb{P}_{\theta_0}$, où θ_0 est inconnu.

À l'instant 0, la loi *a posteriori* est indépendante des observations, et donc

$$\pi_0 = \mu \qquad \text{la loi } a \text{ priori}.$$

À l'instant $n+1$, la loi *a posteriori* se déduit de la loi *a posteriori* π_{n-1} en utilisant la formule de Bayes :

$$\forall \theta \in \Theta \qquad \pi_n(\theta) = \frac{\pi_{n-1}(\theta)\mathbb{P}_\theta(X_n)}{\sum_{\theta\in\Theta} \pi_{n-1}(\alpha)\mathbb{P}_\alpha(X_n)}. \tag{6.2}$$

La formulation précédente de la structure markovienne de l'*a posteriori* est une représentation « création - annihilation » qu'on retrouve dans de nombreuses méthodes stochastiques. On consultera à profit

[DM04] P. Del Moral *Feynman-Kac Formulae*, Springer, New-York, 2004.

Definition 6.2.1 (Identifiabilité) *On dit que le modèle est identifiable si l'application $\theta \mapsto \mathbb{P}_\theta$ est injective. On supposera le modèle identifiable par la suite.*

On peut par exemple quantifier l'identifiabilité du modèle au travers de l'utilisation de distances sur les lois de probabilités. Il en existe un nombre important : distance en variation totale, ditance de Hellinger, divergence de Kullback (ou entropie) pour les plus célèbres. Par la suite, nous formaliserons les choses proprement en utilisant la distance de Hellinger (ce choix est un peu arbitraire et des résultats peuvent également être obtenus en utilisant d'autres métriques).

Definition 6.2.2 (Distance de Hellinger) *Si $\mathbb{P}_1$ et $\mathbb{P}_2$ sont deux lois de probabilités sur un espace E, absolument continues l'une par rapport à l'autre, on définit*

$$d_H^2(\mathbb{P}_1, \mathbb{P}_2) := \int_E \left(\sqrt{d\mathbb{P}_1(x)} - \sqrt{d\mathbb{P}_2(x)}\right)^2 = \int_E \left(\sqrt{\frac{d\mathbb{P}_2(x)}{d\mathbb{P}_1(x)}} - 1\right)^2 d\mathbb{P}_1(x).$$

On pourra vérifier les faits suivants.

Proposition 6.2.1 (Propriétés de la distance de Hellinger) *Les points suivants sont satisfaits :*

i) d_H est une distance (tous les axiomes sont vérifiés).

ii) Elle est d'ailleurs non tributaire de la mesure de référence qui pourrait servir à la définir si μ est une mesure telle que $\mathbb{P}_1$ et $\mathbb{P}_2$ sont absolument continues par rapport à μ.

iii) La distance de Hellinger est toujours majorée par $\sqrt{2}$.

iv) On a la relation avec la distance en variation totale :

$$\frac{1}{2} d_H^2(\mathbb{P}_1, \mathbb{P}_2) \leq d_{VT}(\mathbb{P}_1, \mathbb{P}_2) \leq d_H(\mathbb{P}_1, \mathbb{P}_2)$$

v) On a la relation avec l'entropie :

$$2 d_H^2(\mathbb{P}_1, \mathbb{P}_2) \leq d_{KL}(\mathbb{P}_1, \mathbb{P}_2).$$

Cette distance permet alors de quantifier le résultat de consistance bayésienne lorsque Θ est fini et le modèle identifiable.

Theorem 6.2.1 (Consistance bayésienne) *Supposons que le modèle est identifiable avec*

$$\forall \theta \neq \theta_0 \qquad d_H^2(\theta, \theta_0) \geq h^2 > 0,$$

alors la loi a posteriori se concentre exponentiellement vite vers δ_{θ_0} si $\pi(\theta_0) > 0$.

<u>Preuve :</u> On note π_n l'*a posteriori* à l'étape n. Puisque $\pi(\theta_0) > 0$, on constate immédiatement que $\pi_n(\theta_0) > 0$ pour tout entier n. On peut alors étudier la quantité suivante :

$$\forall \theta \in \Theta \qquad \frac{\pi_n(\theta)}{\pi_n(\theta_0)} = \frac{\pi_{n-1}(\theta)}{\pi_{n-1}(\theta_0)} \times \frac{\mathbb{P}_\theta(X_n)}{\mathbb{P}_{\theta_0}(X_n)}.$$

Si l'on définit le rapport de vraisemblance de la n-ième observation

$$\Lambda_n(\theta) := \frac{\mathbb{P}_\theta(X_n)}{\mathbb{P}_{\theta_0}(X_n)},$$

ainsi que $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ on constate alors que

$$\mathbb{E}\left[\Lambda_n(\theta) | \mathcal{F}_{n-1}\right] = 1.$$

Par conséquent, l'évolution du ratio $\pi_n(\theta)/\pi_n(\theta_0)$ n'est pas totalement informatif (c'est une martingale). En revanche, on peut subtilement considérer la racine carrée de ce ratio :

$$\forall \theta \neq \theta_0 \qquad \mathbb{E}\left[\sqrt{\frac{\pi_n(\theta)}{\pi_n(\theta_0)}}|\mathcal{F}_{n-1}\right] = \sqrt{\frac{\pi_{n-1}(\theta)}{\pi_{n-1}(\theta_0)}} \times \mathbb{E}\left[\sqrt{\Lambda_n(\theta)}|\mathcal{F}_{n-1}\right] < \sqrt{\frac{\pi_{n-1}(\theta)}{\pi_{n-1}(\theta_0)}}$$

puisque l'inégalité de Jensen (cas d'égalité dans l'inégalité de Jensen) implique que

$$\forall \theta \neq \theta_0 \qquad \mathbb{E}\left[\sqrt{\Lambda_n(\theta)}|\mathcal{F}_{n-1}\right] < \sqrt{\mathbb{E}\left[\Lambda_n(\theta)|\mathcal{F}_{n-1}\right]} = 1.$$

On en déduit que $r_n(\theta) = \sqrt{\pi_n(\theta)/\pi_n(\theta_0)}$ est une sur-martingale positive donc convergente. Il est même facile de voir que nécessairement

$$\forall \theta \neq \theta_0 \qquad r_n(\theta) \longrightarrow 0 \quad p.s. \qquad \text{lorsque} \qquad n \longrightarrow +\infty.$$

On conclut donc rapidement en remarquant que par définition de l'*a posteriori* :

$$\pi_n(\theta_0)\left(1 + \sum_{\theta \neq \theta_0} \frac{\pi_n(\theta)}{\pi_n(\theta_0)}\right) = 1.$$

L'ensemble Θ étant fini, on constate que

$$\epsilon_n = \sum_{\theta \neq \theta_0} \frac{\pi_n(\theta)}{\pi_n(\theta_0)} \longrightarrow 0 \quad p.s. \qquad \text{lorsque} \qquad n \longrightarrow +\infty,$$

ce qui implique alors que

$$\pi_n(\theta_0) \longrightarrow 1 \quad p.s. \qquad \text{lorsque} \qquad n \longrightarrow +\infty.$$

On peut être plus précis dans les estimations précédentes. En effet, jusqu'à présent, nous n'avons pas utilisé d'arguments quantifiant la distance à 1 des quantités $\sqrt{\Lambda_n}$. En fait, il est facile de voir que

$$\begin{aligned}
\forall \theta \neq \theta_0 \qquad d_h^2\left(\mathbb{P}_\theta, \mathbb{P}_{\theta_0}\right) &= \int_E \left[\sqrt{\frac{d\mathbb{P}_\theta(x)}{d\mathbb{P}_{\theta_0}(x)}} - 1\right]^2 d\mathbb{P}_{\theta_0}(x) \\
&= \int_E \left(\frac{d\mathbb{P}_\theta(x)}{d\mathbb{P}_{\theta_0}(x)} + 1 - 2\sqrt{\frac{d\mathbb{P}_\theta(x)}{d\mathbb{P}_{\theta_0}(x)}}\right) d\mathbb{P}_{\theta_0}(x) \\
&= 2\left[1 - \mathbb{E}\sqrt{\frac{d\mathbb{P}_\theta(X)}{d\mathbb{P}_{\theta_0}(X)}}\right]
\end{aligned}$$

Ainsi, la racine du rapport de vraisemblance vérifie en réalité :

$$\mathbb{E}\left[\sqrt{\Lambda_n(\theta)}|\mathcal{F}_{n-1}\right] = 1 - \frac{d_H^2(\mathbb{P}_\theta, \mathbb{P}_{\theta_0})}{2} \leq e^{-\frac{d_H^2(\mathbb{P}_\theta, \mathbb{P}_{\theta_0})}{2}}. \tag{6.3}$$

Par conséquent, en utilisant classiquement une méthode de Martingale, on peut définir

$$M_n(\theta) := e^{n\frac{d_H^2(\mathbb{P}_\theta, \mathbb{P}_{\theta_0})}{2}} r_n(\theta),$$

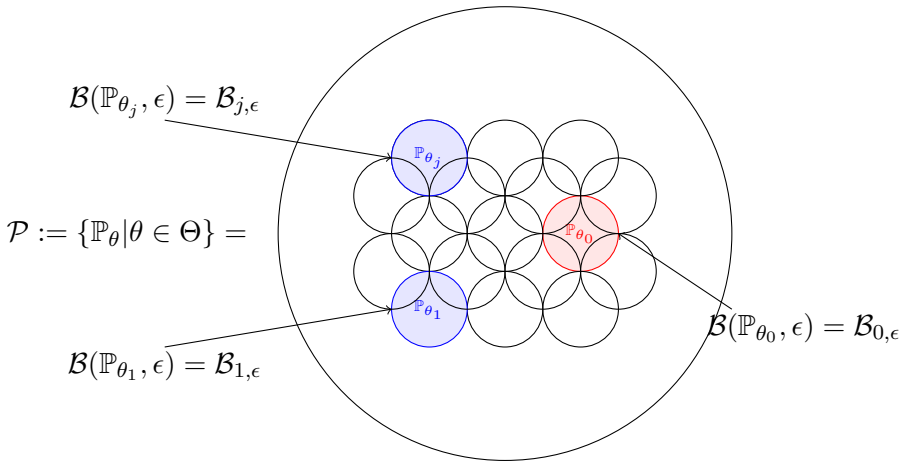

FIGURE 6.1: Représentation de la consistance bayésienne lorsque Θ est recouvert par un nombre fini de boules.

et on constate que $M_n(\theta)$ est une sur-martingale positive, donc convergente p.s. et finalement, il existe une constante C assez grande pour laquelle :

$$\forall \theta \neq \theta_0 \quad \forall n \in \mathbb{N} \qquad r_n(\theta) \leq C e^{-n \frac{d_H^2(\mathbb{P}_\theta, \mathbb{P}_{\theta_0})}{2}}.$$

On peut alors conclure à nouveau en remarquant que

$$\pi_n \left(\Theta \backslash_{\theta_0}\right) \leq C|\Theta| e^{-nh^2}. \tag{6.4}$$

Cela implique *a fortiori* la convergence de π_n vers δ_{θ_0} à vitesse exponentielle. □

6.2.3 Cas où Θ est quelconque

La situation n'est guère plus compliquée en ce qui concerne la philosophie de la preuve précédente. Elle est synthétisée dans la figure 6.1 qui suit.

Étant donné un $\epsilon > 0$, nous allons procéder à un recouvrement de l'ensemble $\mathcal{P}$ par des boules de rayon ϵ **au sens de la distance de Hellinger** Cet argument est en effet incontournable puisque lorsque $\mathbb{P}_\theta$ se rapproche de $\mathbb{P}_{\theta_0}$ (par exemple au sens de Hellinger), il n'est plus possible de « détacher » de 1 la racine du rapport de vraisemblance $\sqrt{\Lambda_n}$ dans l'équation Equation (6.3). Par contre, lorsqu'on considère des θ en dehors de $B(\mathbb{P}_{\theta_0}, \epsilon)$, cette majoration reste valable et peut même être étendue :

$$\mathbb{E}\left[\sqrt{\frac{\int_{B_{j,\epsilon}} \pi_{n+1}(\theta) d\theta}{\int_{B_{0,\epsilon}} \pi_{n+1}(\theta) d\theta}} | \mathcal{F}_n\right] \leq \mathbb{E}\left[\sqrt{\frac{\int_{B_{j,\epsilon}} \pi_n(\theta) d\theta}{\int_{B_{0,\epsilon}} \pi_n(\theta) d\theta}}\right] e^{-c\epsilon^2}$$

Enfin, on conclut la convergence exponentielle du moment que le nombre de boules $(B_{j,\epsilon})_j$ noté $N_\epsilon(\Theta)$ nécessaires pour couvrir Θ par des boules de rayon ϵ (au sens de la distance de Hellinger) n'est pas trop important pour appliquer Equation (6.4) puisque cette inégalité se récrit dans ce contexte :

$$\pi_n \left(B_{0,\epsilon}^c\right) \leq C N_\epsilon(\Theta) e^{-cn\epsilon^2}$$

Enfin, un examen minutieux de la preuve du théorème précédent, et de l'inégalité précédente, permet même d'exhiber une méthode pour quantifier la vitesse de décroissance des boules de rayon ϵ optimale préservant la concentration de π_n dans B_{0,ϵ_n}. On énoncera sans démonstration (laissé en exercice) le résultat suivant :

Theorem 6.2.2 *On suppose que $\mathcal{P}$ peut être recouvert au sens de Hellinger par $N_\epsilon(\Theta)$ boules de rayon ϵ et qu'en plus on a pour une suite ϵ_n tendant vers 0 :*

$$\log N_{\epsilon_n}(\Theta) \leq n\epsilon_n^2, \tag{6.5}$$

$$\pi_0\left(\tilde{B}_{0,\epsilon_n}\right) \geq e^{-Cn\epsilon_n^2}. \tag{6.6}$$

Alors,

$$\pi_n\left(B_{0,\epsilon_n}\right) \longrightarrow 1 \quad p.s. \qquad lorsque \qquad n \longrightarrow +\infty.$$

En réalité, on a besoin d'une minoration du poids *a priori* de la boule B_{0,ϵ_n} car si on part avec un poids trop petit autour de cette boule, la dynamique markovienne n'a pas le temps en n itération d'aumenter assez son poids.

De même, le modèle ne doit pas être trop gros sinon la majoration Equation (6.4) ne dit plus grand chose. On constate donc que dans le cas de modèle paramétrique, la vitesse cible est $\epsilon_n = 1/\sqrt{n}$ tandis que pour des situations en dimension infinie, il faut contrôler finement les dimensions entropiques des modèles. Ce genre de question est plus d'ordre statistique que probabiliste. On consultera à profit l'article [GGvdV00] qui donne un exposé très synthétique de la situation, en relation avec des vitesses *minimax* dans certains modèles. Notez que la preuve qui y est proposée est nettement plus laborieuse que l'utilisation des martingales évoquée plus haut.

6.3 Algorithme EM

L'article fondateur qui peut être consulté est :

[DLR77] A. P. Dempster, N. M. Laird, D. B. Rubin *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society. Series B, Vol. 39, No. 1. 1977.

6.3.1 Contexte

L'algorithme EM *Expectation - Maximisation* est un algorithme d'estimation statistique dans des modèles bayésiens. Il est utile pour trouver l'estimation du maximum *a posteriori*. Cet estimateur est en général noté le MAP pour l'a priori π sur Θ.

Il s'applique dans un contexte où on dispose d'un modèle statistique $(\mathbb{P}_\theta)_{\theta\in\Theta}$ générant des variables aléatoires (X, Z). Le souci d'estimation réside dans le fait qu'on observe seulement la coordonnée X et non le couple (X, Z) alors que la maximisation sur le paramètre θ de la vraisemblance complète n'est possible que si on observe le couple (X, Z).

Ainsi, on définit les objets mathématiques suivant.

Definition 6.3.1 (Vraisemblance, Maximum *a posteriori*) *$(\mathbb{P}_\theta)_{\theta\in\Theta}$ est un modèle générant des variables(X, Z). On suppose observées $(X_1, \ldots, X_n)$ et non observées les variables $(Z_1, \ldots, Z_n)$ sachant que*

$$(X_i, Z_i) i.i.d. \sim \mathbb{P}_{\theta^0}.$$

Étant donné un a priori π sur Θ, l'estimateur du maximum a posteriori est défini par

$$\hat{\theta}_n^{MAP} := \arg\max_{\theta\in\Theta} \{\pi(\theta)\mathbb{P}_\theta\left[(X_1,\ldots,X_n)\right]\} = \arg\max_{\theta\in\Theta} \pi_n(\theta).$$

Si le modèle statistique $(\mathbb{P}_\theta)_{\theta\in\Theta}$ n'est pas trop gros, et que l'*a priori* π charge un voisinage de θ^0 positivement, on sait par le biais des théorèmes de la section précédente que la loi *a posteriori* π_n se concentre sur la valeur θ^0, lorsque le nombre d'observations $n \longrightarrow +\infty$. Il en est donc de même de $\hat{\theta}_n^{MAP}$. Il s'agit donc de trouver un algorithme d'optimisation efficace permettant d'approcher θ_n^{MAP}.

Hypothèse : Nous supposons que la procédure de calcul de la vraisemblance complète est possible, c'est-à-dire la fonction définie par

$$\ell(X,Z,\theta) := \log\left(\mathbb{P}\left[(X,Z)|\theta\right]\right) + \log\left(\pi(\theta)\right),$$

est disponible au travers d'une formule simple.

Rappel de vocabulaire Nous allons décrire un algorithme extrêmement courant pour le calcul du « MAP » dans les situations de données manquantes. C'est une méthode itérative, et il convient de bien distinguer les itérations de l'algorithme $k \in \mathbb{N}$ du nombre de données définissant le « MAP ». Ainsi, on parlera de **convergence** de l'algorithme lorsque $k \longrightarrow +\infty$ et que

$$\theta_k \longrightarrow \hat{\theta}_n^{MAP}.$$

Rappelons que l'estimation, elle, est **consistante** si

$$\hat{\theta}_n^{MAP} \longrightarrow \theta^0 \qquad \text{lorsque} \qquad n \longrightarrow +\infty.$$

Algorithme On définit la suite de paramètres $(\theta_k)_{k\in\mathbb{N}}$ par

— $\theta_0 \in \Theta$ choisi quelconque.

— **Étape E** Étant donné $\theta_k \in \Theta$, on définit l'application

$$\theta \in \Theta \longmapsto Q(\theta_k,\theta) := \mathbb{E}_{Z|X,\theta_k}\left[\ell(X,Z,\theta)\right].$$

Ce calcul correspond à l'étape E car Q est une espérance définie au travers la loi $\mathbb{P}(.|\theta_k,X)$

— **Étape M** On détermine θ_{k+1} comme

$$\theta_{k+1} := \arg\max_{\theta\in\Theta} Q(\theta_k,\theta).$$

Ce calcul correspond à l'étape M car on maximise la fonction $Q(\theta_k,.)$.

Dans la définition de la fonction Q, il faut comprendre que lors de l'itération k, on dispose de la valeur courante θ_k du paramètre, des observations X (issues du n échantillon initial qui lui n'a pas bougé). Les observations Z n'ayant pas été observées, nous les *simulons* au travers de la loi conditionnées aux X observés $\mathbb{P}(.|\theta_k,X)$ et déterminons la valeur de θ_{k+1} comme étant celle qui a maximisé la vraisemblance simulée selon la loi conditionnelle.

Convergence Nous allons démontrer le théorème suivant de croisance de la vraisemblance complète.

Theorem 6.3.1 (Algorithme EM) *La suite $(\theta_k)_{k\in\mathbb{N}}$ définie au travers des itérations EM est telle que $(\log\mathbb{P}(\theta_k|X))_{k\in\mathbb{N}}$ est croissante.*

Preuve : L'idée suit le principe des démonstrations dans les algorithmes « max-max ». On cherche à prouver que $\ell(\theta_k|X)$ est croissante, pour cela, on va minorer par une fonction δ la différence :

$$\ell(\theta|X) - \ell(\theta_k|X) \geq \delta(\theta_k, \theta),$$

et montrer que $\delta(\theta_k, \theta_{k+1})$ est positive. Ainsi, à chaque étape de l'itération courante, on minore la fonction à maximiser par une fonction plus « tractable » qu'on majore.

On passe désormais aux étapes calculatoires :

$$\ell(\theta|X) - \ell(\theta_k|X) = \log \mathbb{P}_\theta(X)\pi(\theta) - \log \mathbb{P}_{\theta_k}(X)\pi(\theta_k),$$

et bien sûr

$$\begin{aligned}
\mathbb{P}_\theta(X)\pi(\theta) &= \sum_Z \mathbb{P}_\theta(X, Z)\pi(\theta) = \sum_Z \mathbb{P}(X|\theta, Z)\mathbb{P}(Z|\theta)\pi(\theta) \\
&= \sum_Z \frac{\mathbb{P}(X|\theta, Z)\mathbb{P}(Z|\theta)\pi(\theta)}{\mathbb{P}(Z|X, \theta_k)} \times \mathbb{P}(Z|X, \theta_k)
\end{aligned}$$

On peut alors remarquer que $\mathbb{P}(.|X, \theta_k)$ est une loi de probabilité et l'inégalité de Jensen s'applique par concavité de la fonciton log. Ainsi

$$\log\left(\mathbb{P}_\theta(X)\pi(\theta)\right) \geq \sum_Z \log\left(\frac{\mathbb{P}(X|\theta, Z)\mathbb{P}(Z|\theta)\pi(\theta)}{\mathbb{P}(Z|X, \theta_k)}\right)\mathbb{P}(Z|X, \theta_k).$$

Par ailleurs, le second membre de la différence à étudier s'écrit simplement :

$$\log\left(\mathbb{P}(X|\theta_k)\pi(\theta_k)\right) = \sum_Z \mathbb{P}(Z|X, \theta_k)\log\left(\mathbb{P}(X|\theta_k)\pi(\theta_k)\right).$$

Par conséquent :

$$\begin{aligned}
\log\left(\mathbb{P}_\theta(X)\pi(\theta)\right) - \log\left(\mathbb{P}_{\theta_k}(X)\pi(\theta_k)\right) &\geq \sum_Z \mathbb{P}(Z|X, \theta_k)\log\left(\frac{\mathbb{P}(X|\theta, Z)\mathbb{P}(Z|\theta)\pi(\theta)}{\mathbb{P}(Z|X, \theta_k)}\right) \\
&\quad - \sum_Z \mathbb{P}(Z|X, \theta_k)\log\left(\mathbb{P}(X|\theta_k)\pi(\theta_k)\right) \\
&= \sum_Z \mathbb{P}(Z|X, \theta_k)\log\left(\frac{\mathbb{P}(X|\theta, Z)\mathbb{P}(Z|\theta)\pi(\theta)}{\mathbb{P}(Z|X, \theta_k)\mathbb{P}(X|\theta_k)\pi(\theta_k)}\right) \\
&= \sum_Z \mathbb{P}(Z|X, \theta_k)\log\left(\frac{\mathbb{P}(X, Z|\theta)\pi(\theta)}{\mathbb{P}(X, Z|\theta_k)\pi(\theta_k)}\right) \\
&:= \delta(\theta_k, \theta).
\end{aligned}$$

On constate alors que le calcul de $Q(\theta_k, \theta)$ est identique à celui de la fonction δ précédente. Ainsi, d'une itération à l'autre, la loi *a posteriori* évaluée en θ_k grandit. Cela finit la preuve du théorème. □

En toute généralité, il est difficile de dire plus que ce résultat car cela réclame alors des propriétés d'unicité du maximum a posteriori. Il existe des critères garantissant ces propriétés d'unicité. On consultera à profit :

[W83] J. Wu *On the convergence properties of the EM algorithm*, Annals of Statistics, Vol 11 (1) 1983.

6.4 Algorithme SA-EM

6.4.1 Motivations

L'algorithme SA-EM est une version « approximation stochastique » de la méthode EM précédente. Une référence pour ce paragraphe, parmi d'autres, est l'article original introduisant cette méthode.

[DLM99] B. Delyon, M. Lavielle, E. Moulines *Convergence of a stochastic approximation version of the EM algorithm*, The Annals of Statistics, Vol 27, 1, 1999.

Il s'agit d'une méthode parmi d'autres introduisant une phase d'approximation stochastique, et on pourra rencontrer également des méthodes Monte-Carlo -EM (MCEM). L'idée principale est de substituer la phase E de la méthode EM par une méthode stochastique remplaçant le calcul de la fonction $Q(\theta_k, \theta)$. On sent que les algorithmes stochastiques vont pouvoir être utiles puisque

$$Q(\theta_k, \theta) = \mathbb{E}_{Z \sim \mathbb{P}_{\theta_k}(.|X)}[\ell(X, Z, \theta)],$$

et que l'algorithme de Robbins-Monro permet de maximiser (ou minimiser) une énergie s'exprimant sous la forme d'une espérance.

6.4.2 Description de l'algorithme

Algorithme Nous donnons ici une version « simplifiée » de SAEM, et les lecteurs intéressés se documenteront dans [DLM99].

- θ_0 initialisé quelconque dans Θ.
- **Étape SA-E :** on simule une donnée « non observée » z_k selon la loi *a posteriori* $\mathbb{P}(.|X, \theta_k)$ où θ_k est la valeur du paramètre de l'itération courante, et X l'ensemble des données observées.
- On calcule alors la nouvelle fonction

$$\hat{Q}_k(\theta) = \hat{Q}_{k-1}(\theta) + \gamma_k \left[\ell(z_k, X, \theta) - \hat{Q}_{k-1}(\theta)\right]$$

- **Étape M :** On trouve θ_{k+1} en maximisant la fonction $\hat{Q}_k$.

Hypothèse de modèle exponentiel Dans un cadre classique de modèle exponentiel, il est possible de rendre les choses un petit peu plus explicites. Nous faisons les hypothèses suivantes.
$(\mathbf{H_M})$: Le modèle est exponentiel et décrit ainsi :

$$\mathbb{P}(X, Z|\theta) = \exp\left(-\psi(\theta) + \langle S(X, Z); \phi(\theta)\rangle\right),$$

où S est une fonction à valeurs dans $\mathbb{R}^l$, $\psi(\theta)$ est une constante de normalisation. Ainsi,

$$\ell(X, Z, \theta) = -\psi(\theta) + \langle S(X, Z); \phi(\theta)\rangle.$$

$(\mathbf{H_{Smooth1}})$: Les fonctions ϕ et ψ sont $\mathcal{C}^2(\Theta)$. La fonction $\theta \longmapsto \bar{s}(\theta) = \mathbb{E}[S(X, Z)|\theta]$ est $\mathcal{C}^1(\Theta)$ ainsi que la fonction $\theta \longmapsto \mathbb{E}\left[\log \mathbb{P}[(X, Z)|\theta]\right] := l(\theta)$. Bien sûr, si on intègre un *a priori* π sur Θ, cet *a priori* s'exprime directement dans la fonction ψ en regardant la loi *a posteriori*.
$(\mathbf{H_{Smooth2}})$: Pour toute valeur de $s = S(X, Z)$, la fonction

$$\hat{\theta}(s) := \arg\max -\psi(\theta) + \langle s, \phi(\theta)\rangle + \log(\pi(\theta)),$$

est une fonction $\mathcal{C}^1(\Theta)$.

Remarquons alors que sous ces hypothèses de modèle exponentiel, tout peut être reparamétré par la valeur de $S(X,Z)$ ou $S(X,z_k)$. Ainsi, si on note la valeur de S à l'étape k par s_k : l'algorithme EM produirait à l'étape E :

$$s_{k+1} = \mathbb{E}_Z \left[S(X,Z) | X, \theta_k \right].$$

L'algorithme SA-EM, quand à lui, donnerait :

$$s_{k+1} = s_k + \gamma_k \left(S(X,z_k) - s_k \right),$$

où $z_k \sim \mathbb{P}[.|\theta_k, X]$. L'étape M, elle, reste inchangée :

$$\theta_{k+1} = \arg\max_{\theta \in \Theta} \left[-\psi(\theta) + \langle s_{k+1}, \phi(\theta) \rangle + \log(\pi(\theta)) \right] = \hat{\theta}(s_{k+1}).$$

La plupart des situations où les algorithmes EM/SA-EM sont appliqués s'écrivent comme des modèles exponentiels, avec des fonctions ϕ et ψ explicites. Cela permet d'alléger le notations et faciliter son implémentation.

6.4.3 Convergence de l'algorithme SA-EM

L'idée de la preuve de la convergence de SA-EM est principalement la vérification des conditions d'applicabilité du théorème de Robbins-Monro, décrit dans le chapitre 3. Rappelons que

$$s_{k+1} = s_k + \gamma_k \left[S(X,z_k) - s_k \right],$$

et

$$\theta_k = \hat{\theta}(s_k).$$

On définit la fonction V :

$$\forall s \in \mathcal{S} \qquad V(s) := -\psi(\hat{\theta}(s)) + \langle s, \phi(\hat{\theta}(s)) \rangle = \max_{\theta \in \Theta} \left[-\psi(\theta) + \langle s, \phi(\theta) \rangle + \log(\pi(\theta)) \right].$$

On établit le résultat en nommant ψ la fonction $\psi + \log \pi$. Sous des hypothèses abrégées, on établit le théorème suivant.

Theorem 6.4.1 *Sous les hypothèses* $(\mathbf{H_M}), (\mathbf{H_{Smooth1}}), (\mathbf{H_{Smooth2}})$ *et la condition sur les pas* $(\gamma_k)_{k \in \mathbb{N}}$ *:*

$$\sum_k \gamma_k = +\infty \qquad \sum_k \gamma_k^2 < \infty.$$

Si la suite $(s_k)_{k \geq 1}$ *est compacte, alors*

$$s_k \longrightarrow \{s \quad | \quad \partial_s V(s) = 0\} \quad p.s. \qquad lorsque \qquad k \longrightarrow +\infty,$$

et bien sûr

$$\theta_k \longrightarrow \{\theta \quad | \quad \partial_\theta l(\theta) = 0\} \quad p.s. \qquad lorsque \qquad k \longrightarrow +\infty.$$

L'idée de la preuve est d'écrire que

$$s_{k+1} = s_k + \gamma_k h(s_k) + \gamma_k e_k,$$

où $h(s_k)$ désigne la direction principale de descente de l'algorithme et e_k est un incrément de martingale. Dans notre situation, on constate clairement que

$$h(s) := \mathbb{E}_Z \left[S(X,Z) | X, \hat{\theta}(s) \right] - s = \bar{s}(\hat{\theta}(s)) - s,$$

et pour appliquer le théorème de Robbins-Monro, nous devons trouver une fonction T telle que $\nabla T, h \geq 0$ pour conclure (voir théorème 3.3.6). Dans notre cas, on définit

$$L(s,\theta) := -\psi(\theta) + \langle s, \phi(\theta)\rangle,$$

et nous savons que

$$\hat{\theta}(s) = \arg\max_{\theta\in\Theta} L(s,\theta),$$

donc

$$-\psi'(\hat{\theta}(s)) + \langle s, \nabla\phi(\hat{\theta}(s))\rangle = 0. \tag{6.7}$$

Cette relation se traduit également comme $\partial_\theta L(s,\hat{\theta}(s)) = 0$ et en différentiant en s, on obtient :

$$D^2_\theta L(s,\hat{\theta}(s))\partial_s\hat{\theta}(s) = -\nabla\phi(\hat{\theta}(s))^t. \tag{6.8}$$

Par ailleurs, en remarquant qu'on peut dériver sous le signe intégral, on a

$$l'(\theta) = -\psi'(\theta) + \langle \mathbb{E}\left[(X,Z)|\theta\right], \nabla\phi(\theta)\rangle = -\psi'(\theta) + \langle \bar{s}(\theta), \nabla\phi(\theta)\rangle$$

En utilisant Equation (6.7) dans la dernière expression, nous obtenons :

$$\begin{aligned} l'(\hat{\theta}(s)) &= \langle -s + \bar{s}(\hat{\theta}(s)), \nabla\phi(\hat{\theta}(s))\rangle \\ &= \langle h(s), \nabla\phi(\hat{\theta}(s))\rangle \\ &= -\langle h(s), \partial_s\hat{\theta}(s) D^2_\theta L(s,\hat{\theta}(s))\rangle \end{aligned}$$

Enfin, on a par chaînage :

$$\begin{aligned} \partial_s\left(l(\hat{\theta}(s))\right) &= l'(\hat{\theta}(s))\partial_s\hat{\theta}(s) \\ &= -\langle h(s), \partial_s\hat{\theta}(s) D^2_\theta L(s,\hat{\theta}(s))\rangle\partial_s\hat{\theta}(s) \end{aligned}$$

On peut alors conclure en calculant que

$$F(s) := \langle \partial_s V(s), h(s)\rangle = \langle -\partial_s\left(l(\hat{\theta}(s))\right), h(s)\rangle = h(s)^t\partial_s\hat{\theta}(s)^t D^2_\theta L(s,\hat{\theta}(s))\partial_s\hat{\theta}(s)h(s) \leq 0.$$

La dernière inégalité est satisfaite car en $\hat{\theta}(s)$, la fonction L est maximale, donc de dérivée seconde négative.

La fonction V est donc décroissante au long des itérations, le théorème de Robbins-Monro permet alors de conclure : la trajectoire converge presque sûrement vers un point critique de V. On obtient alors la conclusion souhaitée pour $(\theta_k)_{k\geq 1}$. □

Remark 6.4.1 *La condition de compacité de la suite $(s_k)_{k\geq 0}$ n'est pas totalement anodine. Elle est parfois évidemment satisfaite lorsque les espaces d'états pour X et Z sont compacts, tout comme Θ et que les densités sont minorés par des quantités strictement positives. Lorsque cette condition n'est pas satisaite, on procède à des méthodes de troncature de la suite $(s_k)_{k\geq 0}$ pour la contraindre à appartement à une suite croissante (au sens de l'inclusion) de compacts. On s'appuie alors sur les méthodes d'approximation stochastique contraintes pour conclure. On pourra consulter sur ces méthodes de troncatures dans les algorithmes stochastiques, outre évidemment [DLM99], le livre :*

[KY03] H. J. Kushner, G. Yin Stochastic Approximation and Recursive Algorithms and Applications. *Springer, New-York, 2003.*

Chapitre 7

Simulated annealing

7.1 Principle of the simulated annealing procedure

7.1.1 Concentration of the Gibbs field

The simulated annealing algorithm is a global minimization scheme for solving the problem

$$\arg\min_{x\in\mathbb{R}^d} U(x).$$

It relies on two principles :

— First, an exploration of the state space through a stochstic process $(X_t)_{t\geq 0}$,

— Second, an exploitation with a cooling schedule.

We define an inverse temperature parameter $\beta > 0$ ($T = \frac{1}{\beta}$ in physics). The Laplace principle shows that the distribution π_β defined by

$$\pi_\beta := \frac{e^{-\beta U(x)}}{\int_{\mathbb{R}^d} e^{-\beta U(y)dy}}$$

concentrates around the *global* minimizer of U when $\beta \longrightarrow +\infty$.

This statement is supported by the next proposition.

Proposition 7.1.1 *Assume U is locally L-Lipschitz, then the distribution π_β satisfies*

$$\pi_\beta\left(U > \min U + \delta\right) \longrightarrow 0 \qquad as \qquad \beta \longrightarrow +\infty$$

Proof : Without loss of generality, we assume that $\min U = 0$. To show this result, we simply write

$$\pi_\beta\left(U > \delta\right) = \frac{\int \mathbf{1}_{U>\delta} d\pi_\beta}{Z_\beta} = \frac{\int \mathbf{1}_{U(x)>\delta} e^{-\beta U(x)} dx}{Z_\beta} \leq \frac{\int \mathbf{1}_{U>\delta} e^{-(\beta-1)U} e^{-U}}{Z_\beta}.$$

Now, we obtain

$$\pi_\beta\left(U > \delta\right) \leq e^{-(\beta-1)\delta} \frac{Z_1}{Z_\beta}.$$

It remains to derive a lower boud of Z_β. We consider L the Lipschitz constant around $\arg\min U$. We can design ϵ_β and B_{ϵ_β} small enough such that

$$\sup_{B_{\epsilon_\beta}} \beta U(y) \leq 1.$$

We then deduce that

$$Z_\beta \geq \int_{B_{\epsilon_\beta}} e^{-\beta U} \geq e^{-1}\lambda\left(B_{\epsilon_\beta}\right)$$

The Lebesgue measure of B_{ϵ_β} is lower bounded by $c\beta^{-d}$ where c is small enough, independent of β. We then obtain

$$\pi_\beta\left(U > \delta\right) \leq c^{-1} Z_1 e^{-(\beta-1)\delta}\beta^d.$$

Hence, $\pi_\beta\left(U > \min U + \delta\right)$ goes to 0 exponentially fast (with β). □

7.1.2 Kolmogorov diffusion

7.1.2.1 Definition of the simulated annealing process

In the previous chapter, we have seen the important influence of the semi-group $(P_t)_{t\geq 0}$ associated to the Kolmogorov-Langevin diffusion :

$$dX_t = -\nabla U(X_t)dt + dB_t. \tag{7.1}$$

In particular, under mild assumptions on U, we have seen that Equation (7.1) defines an elliptic ergodic diffusion whose unique invariant measure is π_2. Moreover, the ergodicity of $(X_t)_{t\geq 0}$ and the rate of convergence of may be translated through a variance convergence and a Poincaré inequality.

We now define the following stochastic differential equation

$$dX_t = -\beta_t \nabla U(X_t)dt + dB_t, \tag{7.2}$$

where $(\beta_t)_{t\geq 0}$ is a smooth increasing and differentiable function that will be set up such that $\beta_t \longrightarrow +\infty$ as $t \longrightarrow +\infty$.

Definition 7.1.1 (Simulated annealing process - generator) *Equation Equation (7.1) defines the simulated annealing process. The associated infinitesimal generator $\mathcal{L}_t$ is*

$$\mathcal{L}_t\varphi = -\beta_t\langle \nabla U(x), \nabla_x\varphi\rangle + \frac{1}{2}\Delta_x\varphi + \partial_t\mathcal{L}_t\varphi$$

It is important above to point out that since the process is inhomogeneous with respect to the time t, we will have to derive some functions of t, although in the classical stationary situation, the functions only depend on the state space x. Nevertheless, if now φ depends on t and x, it is immediate to check that

$$\lim_{h\longrightarrow 0}\frac{\mathbb{E}_{x,t}\varphi(X_{t+h}, t+h) - \varphi(x,t)}{h} = -\beta_t\langle \nabla U(x), \nabla_x\varphi(x,t)\rangle + \frac{1}{2}\Delta_x\varphi(x,t) + \partial_t\mathcal{L}_t\varphi(x,t),$$

when φ is a $\mathcal{C}^\infty(\mathbb{R}^d \times \mathbb{R}_+^*)$ function.

Our goal is to show that when t goes to infinity, the simulated annealing process gets "close enough" to the Gibbs field π_{β_t}. Below, we will use several notations.

Definition 7.1.2 (Distribution of $(X_t)_{t\geq 0}$) *The process being uniformly elliptic, the law of X_t has a $\mathcal{C}^\infty$ density with respect to the Lebesgue measure and is denoted by m_t.*

Definition 7.1.3 (Density with respect to π_{β_t}) *We denote by f_t the Radon-Nikodym derivative of the probability density m_t of the simulated annealing process X_t with respect to the Gibbs measure π_{β_t}, i.e. :*

$$f_t = \frac{\mathrm{d}m_t}{\mathrm{d}\pi_{\beta_t}} \tag{7.3}$$

where m_t is the distribution of $(X_s)_{s\geq 0}$ at time t.

7.1.2.2 Properties of the infinitesimal generator

Using the results obtained in chapter 5, one can see that $\mathbb{R}_+ \ni t \to \mathcal{L}_t$ is continuous and therefore the semi-group $(P_{s,t})_{0\leq s\leq t}$ is smooth . Also by their definition the operators $(P_{s,t})_{0\leq s\leq t}$ are linear and have the following semi-group property : $P_{s,t+h} = P_{s,t} \circ P_{t,t+h}$, for all $0 \leq s < t$ and $h > 0$. Hence, for all $0 \leq s \leq t$, we have :

$$\frac{\mathrm{d}}{\mathrm{d}t} P_{s,t} = P_{s,t}\mathcal{L}_t. \tag{7.4}$$

As will be shown below, bounding the L^2-norm of f_t w.r.t. π_{β_t}, *i.e.*, $\|f_t\|_{\pi_{\beta_t}}$, ensures convergence of the simulated annealing algorithm. However it does not provide enough information about the convergence of m_t to π_{β_t} to deduce a fine convergence rate. This is why we study the evolution of $\|f_t - 1\|_{\pi_{\beta_t}}$ which controls the distance between the two measures. If this quantity is bounded then we obtain the convergence of the simulated annealing algorithm. If moreover it converges to zero, it implies a stronger convergence rate.

Proposition 7.1.2 (Deviations of the simulated annealing algorithm) *Denote by* $J_t = \|f_t - 1\|^2_{\pi_{\beta_t}}$, *then*

$$\mathbb{P}\left(U(X_t) \geq \min U + \delta\right) \leq \epsilon_t + \sqrt{\epsilon_t}\sqrt{J_t},$$

where $\epsilon_t \longrightarrow 0$ *as* $\beta_t \longrightarrow +\infty$.

Proof : We compute

$$\mathbb{P}(U(X_t) > \min U + \delta) = \int \mathbf{1}_{U(x)>\min U+\delta} \mathrm{d}m_t = \int \mathbf{1}_{U(x)>\min U+\delta} f_t \mathrm{d}\pi_{\beta_t}.$$

Then, we write $f_t = (f_t - 1) + 1$ and obtain that

$$\mathbb{P}(U(X_t) > \min U + \delta) \leq \pi_{\beta_t}\left(U > \min U + \delta\right) + \int \mathbf{1}_{U(x)>\min U+\delta}(f_t - 1)\mathrm{d}\pi_{\beta_t}$$

Now, the Cauchy-Schwarz inequality yields

$$\mathbb{P}(U(X_t) > \min U + \delta) \leq \pi_{\beta_t}\left(U > \min U + \delta\right) + \sqrt{\pi_{\beta_t}\left(U > \min U + \delta\right)}\sqrt{J_t}.$$

We obtain the proof with $\epsilon_t = \pi_{\beta_t}\left(U > \min U + \delta\right)$ and Proposition 7.1.1 □

7.2 Convergence of the simulated annealing algorithm in $\mathbb{L}^2(\pi_{\beta_t})$

7.2.1 Differential inequality

In order to prove that the simulated annealing algorithm converges, we will show that

$$J_t \longrightarrow 0 \qquad \text{as} \qquad t \longrightarrow +\infty.$$

To obtain such a result, we will deduce a differential inequality for $J_t = \|f_t - 1\|^2_{\pi_{\beta_t}}$.

Definition 7.2.1 (Notation $\langle U\rangle_{\pi_{\beta_t}}$) *For any time t, we denote by*

$$\langle U\rangle_{\pi_{\beta_t}} := \int U d\pi_\beta$$

the mean of U with respect to π_{β_t}.

One can check that

Proposition 7.2.1

$$\partial_t \pi_{\beta_t}(x) = -\beta_t' \left[U(x) - \langle U \rangle_{\pi_{\beta_t}} \right] \pi_{\beta_t}(x).$$

Proof : We can compute

$$\partial_t \pi_{\beta_t}(x) = -\beta_t' \frac{U(x) e^{-\beta_t U(x)}}{Z_{\beta_t}} + \beta_t' \partial_t \{ Z_{\beta_t} \} \frac{e^{-\beta_t U(x)}}{Z_{\beta_t}^2}.$$

The derivative of Z_{β_t} is easy to handle :

$$\partial_t Z_{\beta_t} = \partial_t \int e^{-\beta_t U} = \int -\beta_t' U e^{-\beta_t U} = -\beta_t' < U >_{\pi_\beta} Z_{\beta_t}$$

Therefore, we obtain that

$$\partial_t \pi_{\beta_t}(x) = \beta_t' \left(< U >_{\pi_\beta} - U(x) \right) \pi_\beta(x),$$

which ends the proof of the proposition. □

According to this result, we then obtain the derivative of I_t.

Proposition 7.2.2 *For any $t > 0$, we have*

$$\partial_t J_t = \partial_t \| f_t - 1 \|_{\pi_{\beta_t}}^2 = 2 < f_t, \mathcal{L}_t f_t >_{\pi_{\beta_t}} + \beta_t' \int (U(x) - \langle U \rangle_{\pi_{\beta_t}}) f_t^2(x) \pi_{\beta_t} dx$$

Proof : We start by computing its derivative :

$$\begin{aligned}
\partial_t J_t = \partial_t \| f_t - 1 \|_{\pi_{\beta_t}}^2 = \partial_t \left\{ \int (f_t^2 - 2 f_t + 1) \pi_{\beta_t} \right\} &= \partial_t \left\{ \int f_t^2 \pi_{\beta_t} \right\} \\
&= 2 \int f_t(x) \partial_t \left[f_t \right](x) \pi_{\beta_t}(x) + \int f_t^2(x) \partial_t \pi_{\beta_t}(x).
\end{aligned}$$

Using the backward Kolmogorov (Fokker-Planck) equation given by Equation (7.4), for the first term we have $\partial_t m_t = \mathcal{L}_t^\star m_t$ where $\mathcal{L}_t^\star$ is the adjoint operator of $\mathcal{L}_t^\star$ in $\mathbb{L}^2(\lambda)$. Therefore, we have

$$\begin{aligned}
\partial_t J_t &= 2 \int f_t(x) \partial_t \left[f_t \right](x) \pi_{\beta_t}(x) dx + \int f_t^2(x) \partial_t \pi_{\beta_t}(x) dx \\
&= 2 \int f_t(x) \frac{\partial_t \left[m_t \right](x)}{\pi_{\beta_t}(x)} \pi_{\beta_t}(x) dx - 2 \int f_t(x) \frac{m_t \partial_t \left[\pi_{\beta_t}(x) \right](x)}{\pi_{\beta_t}^2(x)} \pi_{\beta_t}(x) dx + \int f_t^2(x) \partial_t \pi_{\beta_t}(x) dx \\
&= 2 \int f_t(x) \mathcal{L}_t^\star [m_t](x) dx - 2 \int f_t^2(x) \partial_t \left[\pi_{\beta_t}(x) \right](x) dx + \int f_t^2(x) \partial_t \pi_{\beta_t}(x) dx \\
&= 2 \int \mathcal{L}_t [f_t](x) m_t(x) dx - \int f_t^2(x) \partial_t \left[\pi_{\beta_t}(x) \right](x) dx \\
&= 2 \int f_t(x) \mathcal{L}_t [f_t](x) \pi_{\beta_t}(x) dx + \beta_t' \int f_t^2(x) [U(x) - < U >_{\pi_{\beta_t}}] \pi_{\beta_t}(x) dx.
\end{aligned}$$

Thus, we easily obtain the following equality :

$$\partial_t \| f_t - 1 \|_{\mu_{\beta_t}}^2 = 2 < f_t, \mathcal{L}_t f_t >_{\pi_{\beta_t}} + \beta_t' \int (U(x) - \langle U \rangle_{\pi_{\beta_t}}) f_t^2(x) \pi_{\beta_t} dx, \tag{7.5}$$

which is the desired conclusion. □

7.2.2 Spectral gap asymptotic at low temperature

In the above term, we recover the influence of the Poincaré inequality brought by the first term of the right hand side : $< f, Lf >_{\pi_{\beta_t}} = - < \nabla f, \nabla f >_{\pi_{\beta_t}}$ is expected to be negative and should be related to J_t itself. In order to deal with the first sum we use an estimate of the spectral gap of L_{β_t}. This is provided by Theorem 2.1 of Holley and Strook (see also Freidlin and Wentzell).

Theorem 7.2.1 (Holley and Strook 88) *Under assumptions of* **??**, *there exist two positive constants* $0 < c \leq C < +\infty$ *such that* $\forall\ \beta\ \in\ \mathbb{R}_+$,

$$ce^{-\beta c^\star(U)} \leq \gamma(\beta) \leq Ce^{-\beta c^\star(U)}$$

where $\gamma(\beta) = \inf\{-\int \phi L_\beta \phi\ d\mu_\beta : \|\phi\|_{\mu_\beta} = 1$ *and* $\int \phi d\mu_\beta = 0\}$ *and* $c^\star(U)$ *is the maximum depth of a well containing a local minimum defined in Equation*(7.6).

About the constant $c^\star(U)$ The constant $c^\star(U)$ is always strictly positive as soon as the function has a strict local minimum that is not global. This is generally the case in our setting.

The constant $c^\star(U)$ plays a central role in Theorem **??** and comes from the statement of Theorem 7.2.1. To precisely define $c^\star(U)$, we first introduce some useful notations. For any couple of vertices (x, y) of $\mathbb{R}^p$, and for any path $\gamma_{x,y}$ that connects them, we define $h(\gamma_{x,y})$ as the highest value of U on $\gamma_{x,y}$:

$$h(\gamma_{x,y}) = \max_{s \in \gamma_{x,y}} U(s).$$

We define $H(x, y)$ as the smallest value of $h(\gamma_{x,y})$ obtained for all possible paths from x to y :

$$H(x, y) = \min_{\gamma : x \to y} h(\gamma)$$

Now, for any pair of vertices x and y, the notation $\gamma_{x,y}$ will be reserved for the path that attains the minimum in the definition of $H(x, y)$.

Finally, we introduce the quantity that will mainly determine the size of the spectral gap involved in the functional inequality satisfied by π_β when $\beta \longrightarrow +\infty$ (see the seminal works of Freidlin and Wentzell for an interpretation as a large deviation principle and Holley and Stroock for a functional analysis point of view) :

$$c^\star(U) := \max_{(x,y) \in \mathbb{R}^p} [H(x, y) - U(x) - U(y)] + \min_{x \in \mathbb{R}^p} U(x). \tag{7.6}$$

Figure 7.1 proposes a simplified illustration of the value of $c^\star(U)$ when the state space is of dimension 1. This illustration can be extended to $\mathbb{R}^p$ with the help of a more complex set of possible paths $\gamma_{x,y}$.

For the proof of such a result, we refer to the paper of Holley and Stroock.

7.2.3 Proof of convergence

Below, we assume that $\Delta U < +\infty$, which is a simplification assumption. It is possible to bypass this issue either by considering the process on a compact set, or with a compactification argument (skipped here)

We can continue the study of the simulated annealing process by remarking that Theorem 7.2.1 implies that a constant $c > 0$ exists such that

$$J_t' \leq -ce^{-\beta_t c^\star(U)} J_t + \beta_t' \int (U(x) - \langle U \rangle_{\pi_{\beta_t}})[f_t^2 - 2f_t(x) + 1]\pi_{\beta_t} dx + 2\beta_t' \int (U(x) - \langle U \rangle_{\pi_{\beta_t}})[f_t - 1]\pi_{\beta_t} dx,$$

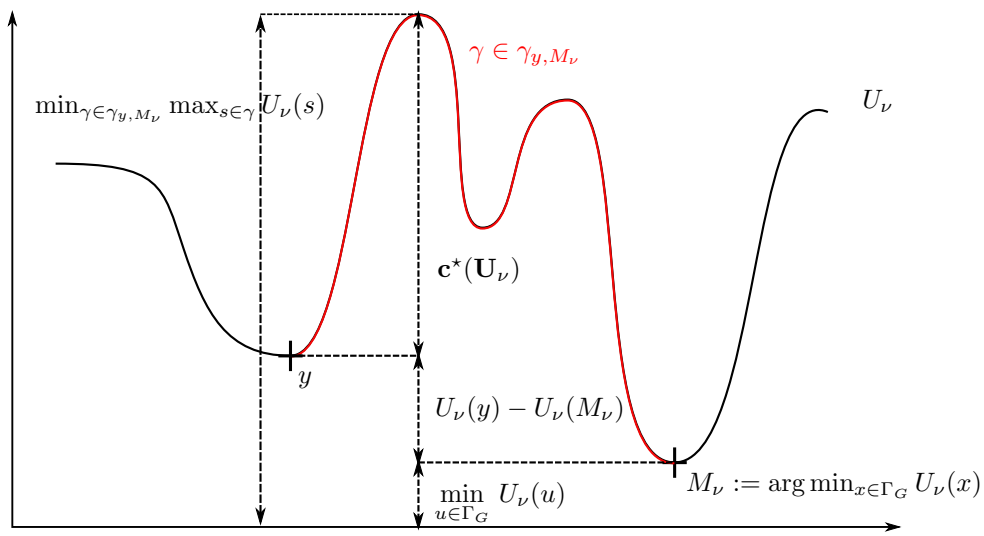

FIGURE 7.1: An example of a function U and of the value $c^\star(U)$ when the possible paths are restricted to horizontal displacements.

which implies with the Cauchy-Schwarz inequality, that

$$J_t' \leq -\left[ce^{-\beta_t c^\star(U)} - \beta_t' \Delta U\right] J_t + M\beta_t' \sqrt{J_t}$$

where M is a sufficiently large constant.

Then, we can choose $\beta_t = \beta \log(1+t)$ and check that as soon as

$$\beta c^\star(U) < 1,$$

then

$$\int_0^{+\infty} e^{-\beta_t c^\star(U)} - \beta_t' dt = +\infty.$$

Then, the Gronwall lemma yields

$$\lim_{t\longrightarrow+\infty} J_t = 0.$$