Habilitation à diriger les recherches

Sébastien Gadat
Institut de Mathématiques
Université de Toulouse (UMR 5219)
31062 Toulouse, Cedex 9, France
Sebastien.Gadat@math.univ-toulouse.fr

7 juin 2012

Table des matières

Chapitre 1

Introduction

In this first chapter, I briefly present several works considered after my PhD defense 6 years ago. All these works are emphasized with respect to the literature. Among this themes, I will just cite some of them and focus on some other contributions with more details in chapters 2, 3 and 4. I will also provide few perspectives for further works.

1.1 High dimensional statistical problems

The study of large dimensional estimation problems is one of the main challenging questions of nowadays statistical works. Let be given labelled observations $((X_1, Y_1, \dots, (X_n, Y_n))$, one important task wishes to predict these labels given a new observation X_{new} without the knowledge of the joint model (X, Y). When X has a low dimensional structure, this question is now somewhat standard although in the opposite case of large dimensional setting, this task is much more difficult. This framework arises in a large number of practical problems such as signal and image processing for instance. They all present the similarity to face the curse of dimensionality which aims to learn the nature of an object which is described with a large number p of features when only few samples n with $n << p$ are available. I briefly list the main ideas of methods historically proposed to overcome the large dimensional difficulty.

1.1.1 State of the art

This problematic has received a great interest during the twenty last years when one consider that Y is a real random variable that must be predicted using a linear combination of features of X. Several approaches have been developed to answer the question of this prediction when p is larger than n. Each of them usually aims to build an estimator $\hat{f}_{n,p}$ that minimises a loss function L which is generally quadratic

$$L(f) = \mathbb{E}[f(X) - Y]^2,$$

where the former expectation is computed with respect to the *unknown* joint law of (X, Y).

Penalized Methods The first historical penalized methods enable to proceed the estimation of f without any real feature selection step and only aim to kill the variance of estimation when one faces a too large dimensional dataset. This is for instance the case when f is estimated by a linear model $f(X) = {}^t\theta X$ penalized by the L^2-norm of the regressor θ. These methods hence solve the minimization of

$$L_{n,p} = \|{}^t\theta X - Y\|_n^2 + p_n(\theta) \tag{1.1}$$

where $\|\|_n$ stands for the empirical norm. When $p_n(\theta) = \lambda_n\|\theta\|_2^2$, we obtain the *Ridge* regression introduced in [Hoerl and Kennard, 1975] that uses a Tikhonov approach (see *e.g.* [Tikhonov, 1943]) and aims to regularize an ill-posed inverse problem, (which is naturally the case when $p >> n$ considering a linear model). This method has enabled to build more sophisticated L^2-Hilbertian estimations using smoothing splines in Reproducing Kernel Hilbert Spaces for instance described in [Wahba, 1990] for instance. The calibration of the penalization term is usually an important step and introduces a bias for small sample size but this bias disappears when n growths to $+\infty$, this is the case when using the AIC or BIC criteria [Akaike, 1974] or the so-called model selection approach of [Barron et al., 1999].

Algorithmic methods It is also natural to refer to methods which limit the overfitting effect in large dimensional setting and are usually inspired from algorithmic ideas such as the CART or Random Forests methods (see [Breiman et al., 1984] and [Breiman, 2001, Amit and Geman, 1997]). In the CART method, the stopping criterion acts as a penalized term in regression or classification to limit the number of leaves built by the algorithm and allows to avoid some overfitting. With algorithms such as Random Forests, it is the randomisation and averaging of uncorrelated predictors which enable to remove overfitting and [Biau et al., 2008] prove that one can obtain consistent procedures using such approaches. This idea of agregating estimators has also been exploited successfully in recent works of Tsybakov when facing classification tasks (see [Tsybakov, 2004] for instance).

Multi-resolution analysis When f is described with a countable family of coefficients, some methods deal with the non parametric estimation of f assuming that the target belongs to a functional space that describes some regularity properties (in general, some Sobolev ir Besov spaces are in force) and use a multi-resolution analysis of the data. The pioneering works of [Donoho et al., 1995, Donoho and Johnstone, 1995] describe some thresholding methods in wavelet basis which enable to limit the number of wavelet coefficients and keep some statistical reconstruction ability that can be adaptive to the nature of the underlying functional space where f lives. Moreover, it is possible to convert these procedures into minimax ones with respect to the quadratic loss (see [Donoho and Johnstone, 1998]). In these approaches, it is thus the assumption on the functional space which permits to solve statistically the estimation of f.

Sparse methods During the last ten years, a large amount of works describe the problem of estimating f with linear predictors computed on observations X using new ideas introduced by *Non-Negative Garotte* method of [Breiman, 1995]. This method has successively inpired the Lasso approach [Efron et al., 2004] for which the main idea is to use the geometric structure of the ℓ^1 ball of dimension p. Indeed, the minimization of (1.1) when $p_n(\theta) \propto \|\theta\|_{\ell^1}$ will yield sparse solutions and obtain naturally feature selection that enable to control the overfitting of the estimation. Moreover, algorithms are available to find such minimizers using some convex analysis tool. Several works use these ideas and one may mention among them the Elastic Net [Zou and Hastie, 2005] which uses a penalization term as a sum of ℓ^1 and ℓ^2 norm, the Dantzig selector [Candes and Tao, 2007] and a lot of generalization of the Lasso ([van de Geer and Bühlmann, 2009, van de Geer, 2008, Bickel et al., 2009] which is a clearly non exhaustive list). The consistency of such procedures is asserted provided that such hypothesis are made on the structure of f. Generally, f is assumed to be s-sparse and even if this assumption is clearly not equivalent to the functional space hypothesis of the former paragraph, it is a structural assumption on the signal to be recovered. Furthermore, the sample size may not be arbitrarily small since in general one may assume that $\log p \sim n$, and one may draw a parallel

with the thresholding methods of the former paragraph that usually keep a number of wavelet coefficients proportional to C^n.

Greedy algorithms At last, some algorithms [DeVore and Temlyakov, 1996] are originated from the approximation theory and are known as *Greedy Algorithms*. These iterative methods in the deterministic case use a general dictionary and build a sequence of approximation of f which are more and more accurate. In the statistical community, these algorithms are called *Boosting* although the approximation theory community refers to *Matching Pursuit* [Davis et al., 1994]. Again, a large amount of recent works exist ([Binev et al., 2005, Donoho et al., 2006, Donoho et al., 2007]) and describe some oracle properties of best approximation using Lebesgue-type inequalities. The principal idea of these methods is to recursively build estimations of the residual between f and its approximation using the *best* predictor in the dictionnary. This idea has been used in a noisy setting in the works of [Bühlmann and Yu, 2003] and such methods associated to a good stopping criterion enable to find a sparse representation of f, even if the dictionary possesses some correlated predictors. At last, one should remark that indeed such method has been also used in the learning theory for classification tasks (see [Freund and Schapire, 1997] for instance) where the original idea was to sequentially build estimations which focus on the hardest samples to be predicted, and then average all the estimations using a suitable stopping criterion.

1.1.2 PhD works on feature selection for supervised classification

In my thesis[1] supervised by Laurent Younes, I worked on the problem of feature selection for supervised classification in a large dimensional setting. Let be given a n sample $(X_1, \dots, X_n)$ described by a large number p of features, we aim to select few meaningful ones. The goal is twofold : improve the ability of classification of the subset of features comparing to the whole set of variables and also understand the meaning of important features. These two objectives are important, one for a natural algorithmic efficiency and the second for the original framework (in genetic for instance, it may be important to understand what are the structuring genes of a biological behaviour).

Many works concern the problem of feature selection for regression task but surprisingly, there exists scarce reflexion about the same goal dedicated to a classification problem. Usually, one can split the existing algorithms in two classes : the first are "filter methods" and îs a preprocessing step before the classification. They thus work with any method of discrimination and are generally using some heuristic criterion to focus on a subset of features. One should consult [Guyon et al., 2006] for a large list of such filter approaches, most of them are not supported by any theoretical justification since they are decorrelated to any classification algorithm and few results are available concerning their consistency. The second approaches are *wrappers* ones and are based on an optimisation step dedicated to a classification algorithm $\mathbb{A}$. One of the main available algorithm is the so-called *Recursive Feature Elimination* [Guyon et al., 2002] which sequentially delete features with poor influence of the margin of classification of a SVM using a backward strategy. One should also consider some recent advances based on a ℓ^1 penalized SVM [Bi et al., 2003, Zhu et al., 2003] exploiting some Lasso ideas.

The method developed in my thesis belongs to the second family of wrapper methods but is slightly different from a method such as the RFE one since the method works with any classification algorithm $\mathbb{A}$. More precisely, if one denote $\mathcal{D}$ the dictionnary of features available on the n sample $X_1, \dots X_n$ labelled by $(Y_1, \dots Y_n)$, and if $\mathbb{A}$ is the supervised algorithm, we aim to mimic a « *best subset* » approach to find a good $\omega \subset \mathcal{D}$. Since numbering all the subsets is numerically

untractable, we have worked on designing a stochastic algorithm which explore not exhaustively some subsets of $\mathcal{D}$. Two theoretical papers have been written on this subject. The first [6] describes the algorithm and provide a complete numerical study in signal processing when $\mathbb{A}$ is a *Support Vector Machine* classifier. The algorithm works as a meta-method and aims to weight features of $\mathcal{D}$ in order to minimize a classification criterion. The method is sequential and some ideas can be compared to the Boosting methods (see [Bühlmann and Yu, 2003] and [Freund and Schapire, 1997]) since the algorithm decrease the weights of some variables proportionately to the classification error observed using a sampled subset of variables. Thus, it can be considered as a boosting algorithm on the feature space.

The second work [5] generalizes the method and proposes to build some tree-structured features with binary composition of elements in $\mathcal{D}$. He is largely inspired from the Random Forest algorithm [Breiman, 2001] and propose a stochastic reversible exploration of forests of binary trees.

In this memory, I have chosen to shortly describe the original feature selection model (Optimal Feature Weighting) as well as the stochastic algorithm developed to solve this model in paragraph 2.1 since several developments have been motivated from this first work.

1.1.3 Biostatistic applications

I have been naturally lead to work on real-data microarrays classification problems after I arrived in Toulouse since some researchers of the *Institut National de Recherche Agronomique* were looking for supervised classification method which also yield dimensionality reduction. With Kim-Anh Lê Cao, we aimed to extend the simulation on the OFW developed in my PhD using different algorithms $\mathbb{A}$ such as CART [2]. We then consider a multi-class framework [3] since it was the natural framework for the INRA-datasets and consider as well the numerical accuracy as the stability results and the biological interpretations of the feature selection method. Even if from a mathematical point of view, this collaboration was restricted to numerical simulations, I consider it fruitful for several reasons described in paragraphs 1.2.3, 1.3 and 1.4.

1.2 Large dimensional estimation problems

I briefly present my works after my PhD on statistical estimation in a large dimensional setting. All these works are concerned with the inference of some rare events comparing to the number of available experiments in the database. The several approaches are all algorithmic but some of them provide also some theoretical developments (paragraphs 1.2.1 and 1.2.3), another one use an extended modelling step (paragraph 1.2.2) and finally the last one is an industrial collaboration with some stochastic methods of rare events simulation (paragraph 1.2.4).

1.2.1 Sequential design of experiments

This collaboration with Serge Cohen and Sébastie Déjean deals with the framework of design of experiments for complex computer models. We can describe the problem as follows : we aim to approach a function f with as few measuring points as possible, to compute an estimation $\hat{f}$ since in this framework, running the code at a design point is generally costly. When the estimation is linear, there exists some almost explicit criterion that quantify the supposed efficiency of the design to compute $\hat{f}$ and these criterion are usually based on the variance of $\hat{f}$. One could refer to pioneering works of [Kiefer and Wolfowitz, 1959, Fedorov, 1972] which provide several optimality criterion for linear models. In our work, we decided to build sequential estimators of f

and we choosed to find the $k+1$-th design point x_{k+1} after the computation of a noisy realisation of $f(x_k)$). Hence, this sequential approach is similar to the method used in [Pronzato, 2000] but we leave open the possibility to control the bias of the model using a minimax approach already given in [Oyet and Wiens, 2000]. Moreover, we propose to use a flexible family of features which randomly vary all along the iteration of the algorithms following a stochastic algorithm. It introduces an alternative approach to the work of [Biswas and Chaudhuri, 2002] which considers a backward testing strategy to obtain a model selection algorithm.

Using a similar strategy of tree exploration described in [5], we have developed in [4] a new stochastic algorithm on a multi-resolution analysis to recursively fix new optimal design points for the estimation of f. Moreover, we have proved a localisation theorem of optimal designs for a particular case of multi-resolution Schauder family which yielded a very fast sequential algorithm. This theoretical result is not obvious since there is from the wavelet nature of the family, no T-systems property is available (see [Dette and Studden, 1997]) for such multi-resolution family. This work is detailed in paragraph refhd :statcomp.

1.2.2 Community graph recovering

With Nathalie Villa, we developed in [7] a graph-clustering algorithm in order to obtain an unsupervised classification method for vertices in a community graph. Usually, a graph $\mathcal{G}$ is given through the definition of its adjacency matrix W which describes the presence of one oriented relation between two vertices. Hence, clustering methods will generally depend only on the structure of W (see for instance the spectral analysis of [Newman, 2006]). However, some general methods may not be adapter to the natural (or expected) partition structure of the graph. Our idea is to exploit some a priori empirical remarks of the community graph structures to build a correct model. Community graphs are known to be structured around clusters where there exists a large number of links between each vertex of the same cluster and oppositely, the presence of one link between two vertices of two different clusters is very unlikely.

Let be given a non oriented symmetric adjacency matrix W with a vanishing diagonal, the degree of each node i is the number of vertices related to i. More precisely, $W_{i,j} = W_{j,i} = 1$ if i and j are linked although $W_{i,j} = W_{j,i} = 0$ in the opposite situation. Of course, the degree of i satisfies $d_i = \sum W_{i,j}$. For any classification $C_1, \dots C_k$ of the set of vertices, the Q-modularity is defined through

$$Q(C_1, \dots, C_k) = \sum_{\ell=1}^{k} \sum_{i,j \in C_\ell} \left[W_{i,j} - \frac{d_i d_j}{2m} \right].$$

We remark that a clustering $C_1, \dots C_k$ posses a large Q-modularity if one has a large number of intra-cluster links. For numerical reasons, the exhaustive search of best partitions is intractable for graphs with a large size. In [7], we designed a stochastic simulated annealing in order to maximise Q. In this work, we also developed a visual representation algorithm to show the obtained partition after the simulated annealing procedure. This step is almost as important as the clustering one in order to obtain a good visualisation of the results. Note that a more recent work [Rossi and Villa, 2010] also exploits the Q-modularity which is maximised with a deterministic simulated annealing *via* a mean field approximation.

1.2.3 Sparse multivariate regression and gene network recovery

My first works with researchers of the *Institut National de Recherche Agronomique* and the former study on graph clustering lead me to work on the problem of the estimation of a network of regulation genes and this field is important to obtain new lightning on biological

processes of genetic diseases. The problem is as follows : two type of datas are computed for a sample of n subjects, the first one E corresponds to the expressions dataset and is a matrix of size $n \times p$ where p is the number of genes considered in the study. E quantifies the amount of expression of each gene on each element of the dataset. The second type of features are discrete marker variables of size $n \times p$.

An interaction between two genes is then described by the fact that one protein activated by one gene acts or inhibits another genes. We describe this interaction using a multivariate linear model

$$E = E\beta + M\alpha + \epsilon, \tag{1.2}$$

where ϵ is the gap between the theoretical interaction and the real process, β is $p \times p$ matrix with vanishing diagonal which provides the structure of the gene network. One aim to recover both α and β and the main difficulty in (1.2) is that the number n of samples is very small comparing to the number of parameters $2p^2 - p$ to predict .

In [17], we first use some penalized regressions to infer $\hat{\alpha}$ and $\hat{\beta}$ (Lasso, Elastic Net and Dantzig selector) and we finally decided to use a multivariate *Boosting* approach. Such methods was already proposed in [Lutz and Bühlmann, 2006] for the multivariate setting but such extension was mainly driven by theoretical proof considerations in order to adapt former results of [DeVore and Temlyakov, 1996] and [Bühlmann, 2006] (for deterministic and noisy situations). Indeed, the method [Lutz and Bühlmann, 2006] does not exploit all the multivariate nature of the data and does not spread in a natural way the effort of the boosting algorithm. In [17], we modify the boosting algorithm in order to obtain a more natural adaptation to the multivariate situation described in (1.2). Comparing to the first work [Bühlmann and Yu, 2003], we introduce a supplementary boosting step in order to choose the coordinate to predict [1]. Thus, we are lead to resume the study of this algorithm first in a deterministic case and then extend it to the noisy realistic situation. This work is described in paragraph 2.3.

1.2.4 Extreme Value Theory and Estimation

At last, I have worked on a very concrete industrial problem for Thales Alenia Space and the Cnes from 2009 to 2011 concerning an estimation for the Egnos-Galileo localization system. The European Spatial Agency requires that Egnos-Galileo provides a localization in a given confidence region and in the opposite case, returns a user alarm. Moreover, the probability that the system does not return an alarm although the object is not in the confidence region should be less than $p = 10^{-7}$ for any period of 150 seconds. Of course, the real historical position from 2006 to 2009 are available as well as the localization provided by Egnos-Galileo.

Thus, the question which may appear to be a rather trivial problem was to estimate the probability of a true positive alarm in order to decide whether this probability is less than $p = 10^{-7}$ or not. Indeed, these events with such weak probability are rarely observed, even if one gets a dataset that furnishes real and localized position during three years and one cannot reduce the estimation to a simple empirical mean.

Our first work [25] uses the Extreme Value Theory described by the so-called Fisher-Tippett law (1928) which asserts that under technical independence conditions, the law of large value of a n samples can be described and depends only on few parameters. More precisely, we have used the Peak Over Threshold (POT) (see *e.g.* [Rassmussen, 1994, de Haan and Ferreira, 2006]) approach to compute an estimation of the true positive alarm for the Egnos-Galileo navigation system. This collaboration with Cécile Mercadier and Jean-Marc Azaïs yields a technical report [37] and a first software.

1. That's why one can consider this as a « boost-boost »algorithm.

A second work has considered such rare events estimation using another point of view which consists in the reinforcement by splitting algorithms the occurences of such feared events (see the stochastic methods described in [Lagnoux-Renaudie, 2009, Lagnoux, 2006]). The principle of such estimation is to use hierarchical duplications of Monte-Carlo simulations in order to generate more and more rare events. We have written a technical report [36] with Agnès Lagnoux, Cécile Mercadier and I and provided it to Thales Alenia Space in order to draw fair comparisons with their initial Petri network approach.

A last work was concerned by the development of an algorithm which yield automatic procedures for the application of Extreme Value Theory approach described in [37]. There were two main difficulties : the first one dealt with the non-stationary nature of the chronological series and was got round by the use of a Portmanteau test. The second difficulty tackled the question of the calibration of a threshold parameter which quantifies when one has a large value of the sample or not. To answer this subtle problem, we have used several algorithms such as [Drees and Kaufmann, 1998, Beirlant et al., 1999] but indeed we found that the more competitive one for this type of datasets were [de Sousa and Michailidis, 2004] and exploits some results on the law of cumulative sums of large values of samples. From this last study, a technical note [35] and a final software has been written by Jean-Marc Azaïs and I, and Thales Alenia Space is currently introducing these tool in the last upgrade of the Egnos-Galileo system.

1.3 Random deformation of signal processing

My initial work on micro-array datasets leads me to the conclusion that most of the time, a good modelling of the structure of the data may largely improves a pure strength algorithm to face a statistical problem. During my PhD, I have worked on handwritten digits recognition problems with Mnist and US Postal database. These data may be considered as a typical example of problem that can be faced using a classification algorithm such as SVM which aims to predict a good object for data that are around a mean expected value, but that can also be considered as a realization of a more complex stochastic process. In this part, we aim to model a stochastic version of signal deformation in order to estimate generative parameters and then improve the signal processing task we can think of.

1.3.1 State of the art

Deformable models When I arrived in Toulouse, Jérémie Bigot has just also complete a work on a statistical method for landmarks registration between noisy images. We naturally start a collaboration around the framework of deformable models. We have tried to use the stuff already available in the deterministic setting and extend it to a noisy case which is a more natural case. We were inspired as well from the works of Alain Trouvé and Laurent Younes on diffeomorphisms built from vector fields and ordinary differential equation as the approach of Grenander [Grenander, 1993a, Grenander and Miller, 2007].

In a general way, the deformation model is described as follows : a mean reference pattern $f^\star$ is defined on $\Omega \subset \mathbb{R}^d$, and we observe some noisy realisations of

$$Y_i(x) = f_i(x) + W_i(x), \quad \forall x \in \Omega, \quad \forall i \in \{1 \dots n\}. \tag{1.3}$$

The variables f_i correspond to a deformation of the reference form $f^\star$ with a random deformation although the W_i represent an additive measurement noise. Moreover, the deformations are assumed to belong to a group G of diffeomorphisms of Ω. Hence,

$$\forall i \in \{1 \dots n\} \quad \exists\, g_i \in G \quad \forall x \in \Omega \qquad f_i(x) = f^\star(g_i.x),$$

where $x \mapsto g_i.x$ refers to the action of g_i on Ω. One can then consider two family of problems. The first one is considering the estimation of the deformation parameters g_i and the second corresponds to the estimation of $f^\star$ itself. My works intensively study this last problem in several situations. Most of the time, G is a finite dimensional Lie group (rigid deformations) or infinite dimensional (elastic deformations). One can immediately remark that a simple empirical averaging which does not take into account the deformation effects, and this cannot reach a satisfactory result as pointed by Figure 1.1.

FIGURE 1.1 – Empirical « naïve » mean between 5 images of a face taken into the Olivetti database [Samaria et al., 1994].

This blurring phenomenon shown by Figure 1.1 reveals that the computation of $f^\star$ as if observations belong to a flat euclidean space is not possible. More precisely, let us define an Hilbert space $\mathcal{H}$ which contains the realisations $(Y_i)_{i\in 1\dots n}$, the empirical mean is defined through the solution of the optimisation problem

$$\bar{Y}_n = \arg\min_{f\in\mathcal{H}} \sum_{i=1}^{n} \|Y_i - f\|^2_{\mathcal{H}}. \tag{1.4}$$

When $\mathcal{H}$ is described by a euclidean distance and when the random deformations $g \in G$ are coming from a law h, $\bar{Y}_n$ is estimating $\tilde{f}$ defined by the convolution

$$\tilde{f}(x) = \mathbb{E}_{g\in G} f^\star(g.x) = \int_G f^\star(g.x)h(g)dg,$$

and of course $\tilde{f} \neq f^\star$. One can then deduce the blurring effect observed with the empirical mean.

It is quite tempting to use some deformation-adapted metrics on $\mathcal{H}$ to compute an estimation with (1.4). This approach is proposed by [Joshi et al., 2004, Miller and Younes, 2001, Trouvé and Younes, 2005] where $\mathcal{H} = L^2(\Omega)$ and the distance is defined by

$$\forall (f_1, f_2) \in \mathcal{H}^2 \qquad d_G(f_1, f_2) = \inf_{g\in G} \left\{ \int_\Omega [f_1(x) - f_2(g.x)]^2 dx + \lambda D(g, e) \right\}, \tag{1.5}$$

where e is the identity of G, λ a regularizing parameter and D a distorsion measure between g and e which quantifies an amount of deformation for the element g.

The use of a non euclidean metric such as the one given by (1.5) to compute $f^\star$ leads to the concept of intrinsic Fréchet mean [Fréchet, 1948] of the distribution as well as the intrinsic Fréchet mean of the n samples. The behaviour of such estimators based on (1.4) with observations that belong to a finite dimension Riemanian manifold are well known, see for instance a complete study in [Bhattacharya and Patrangenaru, 2003, Bhattacharya and Patrangenaru, 2005] for consistency results when $n \mapsto +\infty$. These results are obtained with M-estimation strategy

coupled with riemanian geometry and [Le, 1998, Le and Kume, 2000] have deduced some consistency result for Frechet mean of planar curves dealing with the special situation of the Kendall space of forms [Kendall, 1984].

All these works are largely following a geometric point of view, and does not tackle the natural extension to non parametric estimation for curves and images.

Non parametric statistical approach The estimation of $f^\star$ in the model (1.3) has surprisingly received few attention from a non parametric point of view. Pioneering work of [Kneip and Gasser, 1988] introduces the *shape invariant model* and proposes to approach $f^\star$ when $d = 1$: observations are curves which are parametrized by a known finite number of coefficients and g acts as a translation on $\Omega = \mathbb{R}$:

$$\forall i \in \{1 \dots n\} \quad \exists \tau_i \in G \quad \forall x \in \Omega \qquad dY_i(x) = f^\star(x - \tau_i)dx + dW_i(x).$$

Some other works [Gasser and Kneip, 1992, Gasser and Kneip, 1995] study the semi-parametric case and exploit some ideas which are connected with Fréchet means. The obtained results depend both on the number of observed curves and on the sampling frequency of each observed curve. Methods proposed by[Wang and Gasser, 1997, Ramsay and Li, 2001, Liu and Muller, 2004] consider more general déformations which are not necessarily restricted to translations and tackle the problem the parametrisation of non rigid diffeomorphisms instead of studying an asymptotic reconstruction of $f^\star$ when $n \mapsto +\infty$.

Regarding now the problem of the deformation parameters estimation, [Gamboa et al., 2007b] and [Vimond, 2010] propose some semi-parametric approach to deduce from these estimations an estimator of $f^\star$ when the sampling frequency of each curve (number of points observed for each curve) is arbitrarily large. At last, [Bigot et al., 2010] generalizes this approach to a arbirarily compact Lie groups which model rigid deformations.

A very different approach of [Allassonière et al., 2007] uses a Bayesian point of view to compute an estimation of $f^\star$ from the observations $(Y_i)_{i=1\dots n}$ and [Allassonière et al., 2009] develops a stochastic algorithm based on SAEM in order to find the profile likelihood maximiser.

Most of the above cited works does not study the convergence rates obtained by their procedures and sometimes, even the statistical consistency is unclear (especially for the Bayesian estimators[Allassonière et al., 2007, Allassonière et al., 2009]). It was thus quite natural to study the convergence rates of estimators of $f^\star$.

1.3.2 Randomly shifted curves model

The simplest model of non parametric problem in deformable models is certainly the following one : we observe a set of n curves $(Y_i)_{i\in[1;n]}$ through a white noise model :

$$\forall x \in [0;1], \quad \forall i = 1 \dots n \qquad dY_i(x) = f^\star(x - \tau_i)dx + \sigma dW_i(x), \tag{1.6}$$

where $f^\star$ is the real function to recover which is supposed 1-periodic. The noise level is given by σ and $(W_i)_{i\in 1\dots n}$ are n independent Brownian motions. At last, the random variables $(\tau_i)_{i\in 1\dots n}$ are n translations independent and describe the deformation process. We assume $(\tau_i)_{i\in 1\dots n}$ to be i.i.d. and independent from the $(W_i)_{i\in 1\dots n}$ and we aim to estimate $f^\star$ and understand in what situation the problem is statistically easy or oppositely difficult.

Asymptotic study $(n \mapsto +\infty)$ We first build an estimation of $f^\star$ for the model (1.6) in an asymptotic setting in [9]. We propose an estimation of $f^\star$ using a hard thresholding procedure

in Meyer wavelet basis. The consistency and the convergence rates obtained are rather similar to some phenomenon encountered in statistical inverse problems obtained in deconvolution models[Johnstone et al., 2004, Carroll and Hall, 1988]. Some additional technical difficulties are due to the supplemental random shift τ_i for the thresholding procedure.

Moreover, it is possible to compute the minimax rate of convergence for the L^2 norm when $f^\star$ belongs to a Besov ball $B^s_{p,q}(A)$. The striking point is that the statistical difficulty appears to be the same as the direct deconvolution inverse problem even if we do not observe some realisations of a white noise model on $f^\star \star g$ but each observations corresponds to the same curve $f^\star$ randomly shifted but not convolved by g. The lower bound computation relies on a technical adaptation of the so-called Assouad's Lemma described for instance in [Bretagnolle and Huber, 1979, Has′minskiĭ and Ibragimov, 1990]. Note also that the idea of [Birgé, 1986] which states that lower bound obtained through Assouad's Lemma can also be recovered by the use of Fano's Lemma seems also true here even if one should also consider modify Fano's Lemma (see for instance [Ibragimov and Has′minskiĭ, 1981]), at last, it appears that similar technical difficulties appears to operate the computation of the lower bound using Fano's Lemma.

Oracle approach In [12], we provide a non asymptotic answer to estimate $f^\star$ using the formalism of oracle inequalities. These work relies on the application of the *Unbiased Risk Estimation* method already used in [Cavalier et al., 2002] for general inverse problems. In our framework, the obtained additional term in the oracle inequality depends on the σ^2 (which is rather standard when one use a white noise model) and an additional term which traduces the ill posedness of the inverse problem when using a deconvolution in a Fourier basis. Remark at last that in [12], very similar tools to those used for the study of statistical inverse problems with partially observed operators described for instance in [Cavalier and Raimondo, 2007, Cavalier and Hengartner, 2005].

All these works rely on a somewhat questionable assumption that the law of the random shifts $(\tau_i)_{i\in 1\ldots n}$ is known. It is of course possible to describe an approach which uses the Fréchet mean approach (see [Bhattacharya and Patrangenaru, 2003] for instance), but the theoretical study is much more difficult in the non parametric setting. These works are described in paragraph 3.2.1 of chapter 3.

1.3.3 Estimation of randomly warped images with rigid or elastic deformations

We can extend the model of randomly warped signals by enlarging the structure of deformation sets which act on the unknown signal $f^\star$. It is quite tempting to consider a group G larger than $(\mathbb{R}/\mathbb{Z}, +)$ which is the situation described above, and when $f^\star$ is not yet a curve but an image. G may content for instance translations and rotations. We develop in [8] a second asymptotic study which generalizes the model of randomly shifted curves to the case of a general compact Lie group G for rigid deformations. Our main tool are spectral analysis on Lie groups such as Peter-Weyl theorem and Fourier transform which has been already used in the statistical deconvolution work of [Koo and Kim, 2008, Kim, 1998, Yazici, 2004]. Again, our statistical procedure carry out an optimal minimax rate by studying carefully in Assouad's lemma the likelihood ratios with respect to the size of the Lie group G.

At last, it is also possible to model more complex deformations handling infinite dimensional groups such as large diffeomorphisms group already described in the works of Trouvé and Younes. We propose to generate elastic deformations using a parametrisation of vector fields and consider

the solution at time 1 of a differential flow using these vector fields. We then use standard M-estimation techniques (see *e.g.* [Van der Waart, 1998]) to asymptotically study the estimation of the mean pattern f^*. Some results are provided in[11] as well as optimization methods to compute such estimators. These works are described in paragraph 3.3.5 and 3.2.2.

1.3.4 Constrained regression

In a secondary importance, one can also use the former approach to build monotonic real functions through the differential flows of vector fields in dimension 1. We use this simple remark to build estimators in regression problems where the function is known to be monotone. This problem has received a special importance since numerous practical examples correspond to this a priori information of isotonic regression. The work of [Hall and Huang, 2001, Dette et al., 2006, Dette and Pilz, 2006] consider this problem using a standard kernel estimator projected on the space of monotone functions.

We choose to avoid this projection step since it may introduce some artificial artefacts and the method presented in [10] uses the fact that all strictly non decreasing functions of $[0;1]$ (for instance) may be written as solution at time 1 of some differential equation governed by a time affine vector field. This method is described in 3.1.3.

1.3.5 Intensitu estimation of a randomly shifted counting processes : the case of Poisson processes

After my works on randomly shifted curves, I have been approached by researchers of the *Institut National de la Santé Et de la Recherche Médicale* in order to understand a special process of protein fixation along DNA. Datasets issued from Chip-Seq analysis count the number of cases whenever a protein is fixed at several place of DNA on several chromosomes and biologists have observed that in some case, this fixation may not be so well localized owing to a biological perturbation at the initialization of the fixation process. This yields the researcher to use a convolution by a Gaussian kernel to smooth the data and then a curve alignment to obtain a "mean" profile of the counting process. It would have been tempting to use our approach on randomly shifted curves estimator described by (1.6) to deal with such data. Nevertheless, the nature of the dataset is really different from white noise model and we propose in [16] a model of randomly shifted counting model using Poisson processes with inhomogeneous intensity λ_i. Each λ_i are supposed to be equal to a common intensity λ up to a random shift and this model is largely inspired from (1.6). Note that such problematic appears also in the recent work of [Sansonnet, 2011] where in this situation shifts are observed as well as the mean empirical intensity and one aims to recover λ.

Our approach use intensively concentration properties of [Reynaud-Bourret, 2003] for Poisson processes and our work belongs to the framework of Poissonian inverse problem also studied for instance in [Cavalier and Koo, 2002, Kolaczyk, 1999]. Our estimation still relies on a multiresolution analysis and we are able to build a minimax estimator using a suitable thresholding procedure. Again, the main difficulty already, encountered in [9], is to obtain a suitable lower bound of estimation that makes appearing the inverse problem nature of the model. Moreover, the theoretical adaptivity to the functional space where λ lives requires non trivial extension of the thresholding procedures used in [9]. This works is briefly described in paragraph3.5.3.

1.4 Irreversible optimisation algorithms

The motivation of these works come from a (strange ?) modification of standard stochastic gradient algorithm by Kim-Anh Lê Cao during the numerical studies described in [2] and [3].

The original stochastique gradient algorithm can be written as follows :

$$\forall k \geqslant 0 \qquad X_{k+1} = X_k + \gamma_k d_k + \sqrt{\gamma_k}\zeta_k, \tag{1.7}$$

where X_k stands for the position of the algorithm at iteration k, γ_k is the algorithm step and d_k is the random direction of descent. These algorithms are commonly used in stochastic control, signal and image processing, game theory or Bayesian estimation ... Under technical conditions on d_k and γ_k which should be sufficiently slowly decreasing, one may show the following (informal) properties.

- If $\zeta_k = 0$, classical martingale tools (see *e.g.* [Duflo, 1997, Kushner and Yin, 2003]) show that the behaviour of $(X_k)_{k\geqslant 0}$ is similar to the discretisation of the ordinary differential equation (up to a suitable time modification) :

$$dX_t = -\nabla U(X_t)dt.$$

- When ζ_k is a random Gaussian perturbation, the former result is no longer true and the algorithm is a diffusion approximation due to the presence of $\sqrt{\gamma_k}\zeta_k$ and as soon as $\mathbb{E}[d_k|\mathcal{F}_k] = -\nabla U(X_k)$, $(X_k)_{k\geqslant 0}$ is a discretisation of the stochastic differential equation :

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

  A non exhaustive bibliography can be found in [Benveniste et al., 1990] or [Benaim, 1996] for a more« dynamical »description of this approximation.

The numerical modification used in [2] was to build a stochastic algorithm which is not Markov :

$$\forall k \geqslant 0 \qquad \tilde{X}_{k+1} = \tilde{X}_k + \gamma_k \frac{\sum_{j\leqslant k}\beta_j d_j}{\sum_{j\leqslant k}\beta_j} + \sqrt{\gamma_k}\zeta_k. \tag{1.8}$$

Such numerical scheme is strongly linked to

$$\dot{x}(t) = -\int_0^t r(s,t)D(x(s))ds.$$

provided technical conditions on d_k, D and γ_k. The following several studies has been motivated by optimisation procedures based on this last differential equation, ordinary or stochastic ones.

1.4.1 Averaged memory differential equation

Past works My first work on this theme has considered the family of differential equations which should be the limit of (1.8). The limiting differential equation has then been written with a« memory gradient » :

$$\dot{x}(t) = -\left(\frac{1}{k(t)}\int_0^t h(s)\nabla U(x(s))ds\right)dt, \tag{1.9}$$

where U is a coercive potential defined on $\mathbb{R}^d$. It is possible to rely this equation with second order differential equation with damping using a suitable time parametrization (detailed in [Cabot, 2009]) :

$$\ddot{y}(t) + a(t)\dot{y}(t) + \nabla U(y(t)) = 0, \tag{1.10}$$

where $y = x \circ \tau$, and τ is solution of $\dot{\tau}^2 = k(\tau)/h(\tau)$. The damping effect is $a = \frac{\dot{k}h+k\dot{h}}{2k^{1/2}h^{3/2}} \circ \tau$. On the second order form, (1.10) the differential equation generalizes several known equations. Among them, the first most famous one is the Bessel equation for the special case $a(t) = 1/t$ and $U(x) = x^2$ whose solutions are proportional to J_0 up to a suitable initialisation condition. We then obtain the asymptotic behaviour $x(t) \sim Ct^{-1/4}\cos(2\sqrt{t} - \pi/4)$.

In the convex optimization community, special cases of such euqations was already known and studied when a is a positive constant. In such case, one recovers the *Heavy Ball with Friction* system described in [Polyak, 1987] and [Antipin, 1994] which already study the optimizing properties of such trajectories. This study has then been extended to a general framework of dissipative equations by[Hale, 1988, Haraux, 1991] : they show that such dynamical systems with constant damping converge towards some critical points of U under technical conditions such as analytic or convex for very large x properties. At last, [Ben Hassen and Haraux, 2011] and [Haraux, 2007] use some damping linked with $\ddot{y}$ in order to improve such optimization properties since adapting this damping to the position and speed of the particle $x(t)$ may be of interest.

Contributions We describe in [13] very precisely the behaviour of our damped second order equation when the time t becomes arbitrarily large for equations (1.9) or (1.10), as well as the behaviour of $U(x(t))_{t\geqslant 0}$. Our main assumptions is the convexity of U for large x and the empty interior of the set of critical points of U. Moreover, we prove some one-dimensional result which are not easily transposable to larger dimensions. At last, we study in [14] some more pathological situation where U possesses some flat part (non empty interior of the set of critical points). We provide some details on this « unordinary » differential equation in paragraph4.1.

1.4.2 Memory diffusion

Link with reinforced stochastic process The stochastic algorithm (1.8) when ζ_k is a Gaussian random variable is a numeric approximation of the stochastic process

$$dX_t = -\left(\frac{1}{k(t)}\int_0^t \dot{k}(s)\nabla U(x(s))ds\right)dt + \sigma dB_t. \tag{1.11}$$

since the stochastic algorithm is corrupted by a Brownian increment $\sqrt{\gamma_k}d\zeta_k$. It is thus natural to study such stochastic differential equation (1.11).

The main difficulty in (1.11) comes from its non Markov nature since the process interacts with all its past through the time averaging of $\nabla U(x_s), 0 \leqslant s \leqslant t$. Thus, such process belongs to the large informal class of self-interacting diffusion. First historical example was introduced by [Coppersmith and Diaconis, 1987] for random walks and then extensively studied by[Pemantle, 1992] for the description of the dynamic of Brownian polymer, see also [Cranston and Le Jan, 1995] for a description of such type of continuous processes.

Among the continuous time processes, self-interacting ones are generally coming from a convolution between a drift functional and the occupation measure which may be normalised (see for instance the work of [Benaïm et al., 2002]) or not (see *e.g.*[Durrett and Rogers, 1992]). The drift term at time t is usually an averaging process which is computed from the several values of $(X_t - X_s)_{0\leqslant s\leqslant t}$. From a technical point of view, [Benaïm et al., 2002] makes an extensive use of asymptotic pseudo-trajectory of random dynamical system first introduced by[Benaïm and Hirsh, 1996]. In some sense, such study should have been possible in our framework even if the situation in[Benaïm et al., 2002] is compact although the process (1.11) may explore $\mathbb{R}^d$. At last, we should also refer to recent works of [Kurtzman, 2009] that deal with non

compact manifolds by a supplementary addition of a confining non-interactive potential in the drift term.

Links with hypo-elliptic processes In[15], we propose to study the process (1.11) by a space enlargement method ot obtain a Markov process. The price to pay is then the necessity to handle a strong degeneracy of the random system on the "enlarged" coordinate. Let us denote $(Y_t)_{t\geqslant 0}$ the process given by the drift in (1.11) at time t, if we set $r = \dot{k}/k$, we then obtain the equivalent coupled evolution :

$$\begin{cases} dX_t = -Y_t dt + \sigma dB_t. \\ dY_t = r(t)(\nabla U(X_t) - Y_t)dt. \end{cases} \tag{1.12}$$

Such equations (1.12) then fall into the framework of hypo-elliptic processes. A large number of theoretical advances occurred this last years, among them one should refer to those of [Helffer and Nier, 2005] or [Villani, 2009] which are interested into the evolution of such evolutions for large time t.

One of the main difficulty for the study of convergence to steady regimes of hypo-elliptic evolutions is the lack of classical functional inequalities for instance associated to the Γ_2 criterion [Bakry and Émery, 1985]. One famous example of such situation is the evolution guided by the Fokker-Planck kinetic equations which has received a large amount of interest as attested by the large number of references on the subject, *e.g.* [Risken, 1989, Eckmann and Hairer, 2003, Hérau and Nier, 2004]) tackle this problem by studying carefully the spectrum of the underlying operator although other works ([Desvillettes and Villani, 2001, Dolbeault et al., 2009]) build some coercive norms which stand for Lyapunov function of the dynamical system in order to use a Gronwall lemma. At last, note that Lyapunov functions should be considered as a powerful [2] since a strong link between the existence of such functions and functional inequalities has been underlined in [Bakry et al., 2008] even if such approach is just an intermediary step to obtain convergence to steady regimes for hypo-elliptic systems.

At last, the hypo-elliptic framework introduces additional difficulty which mainly concerns the existence and regularity of $P_t(z_0, .)$ where z_0 is the initializing point of the process at time $t = 0$. The answers are generally given by the use of Hormander works and his famous sum of squares theorem. From the pioneering works of [Hörmander, 1967], we can find lots of theoretical advances that come from partial differential equation such as the works of [Kohn, 1978, Trèves, 1980], or from Malliavin calculus (see *e.g.* [Kusuoka and Stroock, 1987, Cattiaux, 1992, Hairer, 2011]).

In a similar way, under controllability results, it is possible to obtain some sharp estimations of $P_t(z_0, .)$ using Malliavin calculus as pointed by [Delarue and Menozzi, 2010, Bally and Kohatsu-Higa, 2010] or functional Harnack inequalities (see *e.g.*[Pascucci and Polidoro, 2006, Polidoro, 1997]). Of course, these controllability assumptions are not so much surprising since they are already necessary to obtain some positivity result for the semi-group using the Support theorem of [Stroock and Varadhan, 1972]. One should also refer to [Ben Arous and Léandre, 1991] which states a necessarily and sufficient condition under an assumption of boundedness of the drift coefficients for such positivity.

Contributions In our work [15], we provide some stability result for average gradient diffusion systems which are described by equations (1.12). Under rather technical assumptions on U (U should mainly be convex for large $|x|$ with a growing assumption $U(x)/|x| \to +\infty$), we show that

2. The most one ?

the asymptotic behaviour of the process defined through (1.12) relies principally on the long time behaviour of $r(t) = \dot{k}(t)/k(t)$. In particular, we prove the stability of such process when the memory of the process is not too long, and oppositely that the process should explode when the memory is too large.

Main difficulties concern first the hypo-ellipticity of (1.12) (thus its controllability), and also its stability which relies on the construction of a non-trivial Lyapunov function that enable to bound the processus both in position and speed. At last, it is possible to obtain convergence rates in total variation of the occupation measures invoking Lyapunov type argument associated to regularity estimates of the semi-group and using the approach developed by[Down et al., 1995]. These rates are quite explicit thanks to the recent works of [Douc et al., 2009]. I will detail in 4.2 the study of equations (1.11)-(1.12).

1.4.3 Link with kinetic Fokker-Planck equations

The averaged gradien diffusion written on the coupled form (1.12) is from an aesthetic point of view rather similar to the Fokker-Planck kinetic equation

$$\begin{cases} dX_t = V_t dt. \\ dV_t = (-\nabla U(X_t) - V_t)dt + \sigma dB_t dt. \end{cases} \tag{1.13}$$

There is yet a significant difference between two such processes since concerning Fokker-Planck kinetic processes (1.13), the stationary measure is explicitely known although no formula is available for the averaged gradient system (except in que quadratic case $U(x) = a|x|^2$). In such particular case (1.12) is a Gaussian process and one can easily identify its stationary measure. Nevertheless, equations (1.12) and (1.13) do not seem to be equivalent at one glance.

Since the deterministic process (1.9) possesses interesting optimizing properties and that is why we were lead to study a noisy version by a Gaussian noise (which will be arbitrarily small in the sequel). More thant the stability of the averaged gradient system, the exact computation of the L^2 norm can be very instructive to understand whether if (1.12) can be compared positively to other stochastic optimisation methods.

Contributions In [15], we were able to give only partial responses on the convergence rate to steady regime (only rate within total variation distance are obtained), thus we study in [19] the exact computation of the $\mathbb{L}^2$ norm and of the spectrum of the kinetic Fokker-Planck operator which describes (1.13) since these computations are a little bit easier than those concerning (1.12). We first compute exactly the $\mathbb{L}^2$ norm in special case of potential U in the simplest case $U = ax^2/2$, and $U = 0$ on the torus $\mathbb{T} = [0; 1]$ for processes described by (1.13). Our approach is different from the one used in [Dolbeault et al., 2009] or [Villani, 2009] which use a different decomposition of the kinetic Fokker-Planck operator Such results are detailed in paragraph 4.3.

1.4.4 Averaged diffusion with small parameter

The exact computation of $\mathbb{L}^2$ norm in the above paragraph is not so innocent since our objective is indeed to develop an optimization algorithm based on the averaged gradient system to optimise U. This optimisation could be deduced from a simulated annealing using either (1.12) or (1.13) by letting $\sigma(t) \longmapsto 0$ as $t \longmapsto +\infty$. In the sequel, since σ will become small, we will change our notation and denote him $\sigma = \epsilon$ to stress the small size of the diffusion parameter.

Simulated annealing algorithm Concerning the standard elliptic diffusion in $\mathbb{R}^n$:

$$dX_t = \sqrt{\epsilon(t)}dB_t - \nabla U(X_t)dt, \tag{1.14}$$

it is well known that such process can achieve a global minimization of (see [Miclo, 1992] for instance) provided $\epsilon(t) \longmapsto 0$ with a suitable rate. The efficiency of such algorithm depends both on

1. the convergence rate of $P_t(z_0, .)$ towards its steady regime μ_ϵ when ϵ is constant
2. the convergente rate of μ_ϵ towards μ_∞ when $\epsilon \mapsto 0$.

Especially, this balance between these two convergence rates enable to find an optimal decreasing rate for the simulated annealing process (the more the process converges rapidly to its steady regime when ϵ is constant, the more we can fast decrease $\epsilon(t) \longmapsto 0$ and the best is the algorithm).

More precisely, when ϵ is constant, one can expect in the diffusive elliptic case that the process converges exponentially fast to the steady regime μ_ϵ so that

$$Var_{\mu_\epsilon}(P_t^\epsilon(f) - \mu_\epsilon(f)) \leqslant \exp(-A(\epsilon)t)Var(\mu_\epsilon(f)), \tag{1.15}$$

where $A(\epsilon)$ plays a key role in the calibration of the temperature scheme $t \mapsto \epsilon(t)$. Indeed, [Miclo, 1992, Chiang et al., 1987, Royer, 1989] show that there exists an optimal $d^* > 0$ such that $\epsilon(t) = c/\ln(t)$ ensures the convergence of the simulated annealing (1.14) when $c > d^*$ towards the global minimum $\min_{\mathbb{R}^n} U$. This constant d^* corresponds to the elevation of U (see [Miclo, 1992] for a precise definition of d^*) but is not known during practical simulations. Hence, it is necessary to obtain a estimation of d^* for which admissible temperature schemes will be buit, thus the calibration of a sufficiently large $A(\epsilon)$ is important for the simulated annealing procedure.

Considering now the approach of [Bakry et al., 2008], it is proved that as soon as some appropriate Lyapunov function exists, one can find a Poincaré inequality with a constant C_P which is not optimal [3]. Moreover, convergence rates can bounbed by

$$Var_{\mu_\epsilon}(P_t^\epsilon(f) - \mu_\epsilon(f)) \leqslant \exp(-2/C_P t)Var(\mu_\epsilon(f)).$$

Such approach can provide an admissible valule for the simulated annealing but this value seems to be clearly not optimal since C_P is too large. At last, it would be possible to directly study the asymptotic behaviour of the spectrum of the Markov operator L_ϵ that describes the evolution of (1.14) for small values of parameter ϵ (see for instance section 7 of chapter 6 in [Freidlin and Wentzell, 1984] for compact manifolds that contains a stable equilibrium of the dynamical system). Of course, when $\epsilon \longmapsto 0$, the smallest eigenvalue of $-L_\epsilon$ behaves as $\exp(-\Delta V/\epsilon^2)$ where ΔV is an explicit constantthat depends on the quasi-potential deduced from the large deviations of (1.14). But again, this constant is not accessible from a practical point of view since it requires the whole knowledge of U and the calibration of $\epsilon(t)$ cannot be deduced online from this approach.

Second order models Rather than try to estimate almost unsuccessfully the former constant $A(\epsilon)$ (also denoted $d^\star$ in some works) in inequality (1.15), it is possible to think about some other models instead of first order classical ones such as (1.14) and build another diffusion which naturally converges to steady regime faster, and thus for which the constant $\tilde{A}(\epsilon)$ is certainly greater.

3. in general, C_P is too large

It is tempting to use second oder models since they may possess larger ability to explore the state space (phénomenon already observed in the deterministic setting for the dynamical systme (1.9)). Moreover, [Diaconis et al., 2010b] have proved that one can reach with second order Markov chains better convergence rates to steady regime using a non symetric evolution, which is also the case in Fokker-Planck equations and averaged gradient system. In [20], we study the behaviour of averaged gradient diffusion with small parameter. The first step is to identity a clear asymptotic of the invariant measure ν_ϵ of (1.11) when ϵ becomes small since this behaviour guide the convergence of the process (1.11) to global minima of U. [4]

In[20], we study the case of the second order model (1.11) which is restricted to be homogeneous, which is the simplest case, with memory maps $h(t) = k(t) = e^{\lambda t}$ and we obtain a Large Deviation Principle when $\epsilon \to 0$ for $(\nu_\epsilon)_{\epsilon \geqslant 0}$. Except the quadratic case $U(x) = ax^2/x$, there is no explicit formula of ν_ϵ and thus the quasi-potential which should derive from the Large Deviation Principle may be unclear. We provide in[20] sufficiant conditions on the potential U (and certainly not optimal ones) for which ν_ϵ is concentrated on global minimum of U. These works are described in paragraph 4.4.

4. In a sense, it would have been much more simpler to study the kinetic Fokker Planck equations since the stationary measure associated to (1.13) is explicit $m_\epsilon(x,v) \propto e^{-[U(x)+v^2/2]/\epsilon^2}$ and when $\epsilon \longmapsto 0$, the behaviour of the marginal on x is obvious using the Laplace.

Chapitre 2

Statistical modelling and high dimensional estimation

In this chapter, we describe some advances on estimation problems when samples $(X_i, Y_i)_{i=1...n}$ are available, where each X_i is described by p features which form a dictionnary $\mathcal{D} = (g_1, \ldots, g_p)$. Random variables Y_i are either a label of the class in which X_i is living for classification task, or simply an element of $\mathbb{R}^d$ for regression problems. Our study will handle the case $p >> n$ for which standard estimation methods are not efficient owing to the curse of dimensionality.

2.1 Stochastic algorithm for feature selections

In the supervised classification framework, we consider any classification algorithm denoted $\mathbb{A}$ in the sequel and we aim to find a best subset of features, *i.e.* $\mathcal{G} \subset \mathcal{D}$, such that the predictive power of $\mathbb{A}$ using features of $\mathcal{G}$ is « optimal ».

2.1.1 Model description

We will denote $\hat{\mathbb{A}}_{\mathcal{G},n}$ the classification produced by the algorithm $\mathbb{A}$ using samples $(X_i, Y_i)_{i=1...n}$ and active variables $\mathcal{G}$. The prediction error of $\hat{\mathbb{A}}_{\mathcal{G},n}$ is

$$q(\hat{\mathbb{A}}_{\mathcal{G},n}) = \mathbb{P}_{(X,Y)}[\hat{\mathbb{A}}_{\mathcal{G},n} \neq Y],$$

and an ideal approach would be to select

$$\mathcal{G}^\star = \arg\min_{\mathcal{G}\subset\mathcal{D}} q(\hat{\mathbb{A}}_{\mathcal{G},n}). \tag{2.1}$$

Of course, such optimization (2.1) is numerically $\mathcal{D}$. Moreover, the joint law (X, Y) is unknown and it is impossible to exactly recover q, only estimations using the training set are available and we will denote such estimation$\hat{q}$ (see [6] for more details on the bootstrap strategy to compute such estimation).

Our suboptimal approach to study (2.1) is to weight each elements of $\mathcal{D}$ using a discrete probability $\mathbb{P}$ which will be the object to recover. Let us denote an integer k smaller than n, for a given $\mathbb{P}$ we define a mean energy that quantifies the error of $\mathbb{A}$ when features are sampled according to $\mathbb{P}$:

$$\mathcal{E}(\mathbb{P}) = \sum_{\mathcal{G}\in\mathcal{D}^k} \hat{q}(\hat{\mathbb{A}}_{\mathcal{G},n})\mathbb{P}^{\otimes k}(\mathcal{G}). \tag{2.2}$$

When $\mathbb{P}$ is close to minimizer of $\mathcal{E}$, the discrete probability put large weights on features in $\mathcal{D}$ which enable $\mathbb{A}$ to perform well and thus these features are meaningful for the classification

task. It is thus natural to attempt to minimize $\mathcal{E}$ with respect to $\mathbb{P}$. Remark that indeed $\mathcal{E}$ is a k-th order polynomial on variable $\mathbb{P}$ but its coefficient are unknown.

2.1.2 Gradien descent algorithm

We propose in[6] to minimize $\mathcal{E}$ using a sequential strategy of gradient descent, which may be perturbed by a small diffusive term.The general scheme is described by Figure2.1.

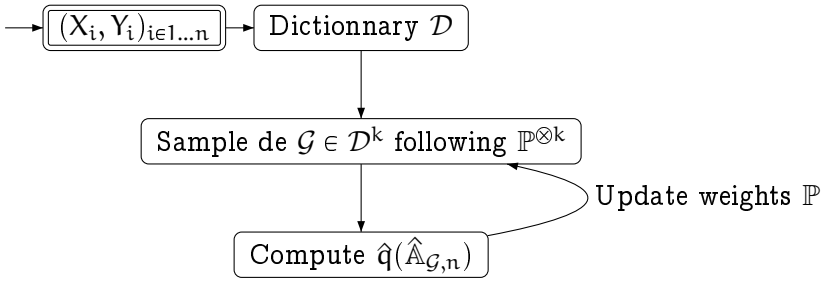


FIGURE 2.1 – Iterative scheme to learn $\mathbb{P}$.

For any point $\mathbb{P}$ that belongs to the simplex $\mathcal{S}_\mathcal{D}$ of discrete probability measures on $\mathcal{D}$, one can compute the euclidean gradient of $\mathcal{E}$:

$$\forall g \in \mathcal{D} \qquad \nabla\mathcal{E}(\mathbb{P})(g) = \sum_{\mathcal{G}\in\mathcal{F}^k} \frac{C(\mathcal{G}, g)\mathbb{P}^{\otimes k}(\mathcal{G})}{\mathbb{P}(g)} \hat{q}(\hat{\mathbb{A}}_{\mathcal{G},n}), \tag{2.3}$$

where $C(\mathcal{G}, g)$ is the number of occurrences of g in $\mathcal{G}$. If we denote $\pi_\mathcal{D}$ the projection on the supporting hyperplane $\mathcal{H}_\mathcal{D}$ of $\mathcal{S}_\mathcal{D}$, such optimization algorithm $\mathcal{E}$ may be as follows :

$$d\mathbb{P}_t = -\pi_\mathcal{D}\left(\nabla\mathcal{E}(\mathbb{P}_t)\right) dt. \tag{2.4}$$

The small parameter diffusion associated to this gradient descent would be defined as the following constrained stochastic differential equation

$$d\mathbb{P}_t = -\pi_\mathcal{D}\left(\nabla\mathcal{E}(\mathbb{P}_t)\right) dt + \sigma_\mathcal{D} dB_t + dZ_t. \tag{2.5}$$

We will denote $\pi_\mathcal{S}$ the projection on the simplex $\mathcal{S}_\mathcal{D}$ since this projection is necessary to build our learning algorithm. In equation (2.5), $(B_t)_{t\geqslant 0}$ is a Brownian motion on $\mathbb{R}^p$ whose covariance matrix $\sigma_\mathcal{D}$ is defined through the projection on $\mathcal{H}_\mathcal{D}$ and dZ_t is a jump process which constrains the process $(\mathbb{P}_t)_{t\geqslant 0}$ to be a discrete probability distribution on $\mathcal{D}$. We won't provide enough technical details to properly state existence and uniqueness of solutions of (2.5). These results are given in [5] and intensively use the Skorokhod map described for instance in[Dupuis and Ramanan, 1999].

2.1.3 Stochastic gradient approximation

Our idea is to use only one computation of $\hat{q}(\hat{\mathbb{A}}_{\mathcal{G},n})$ at each step of the algorithm, this point is not so obvious since equation (2.3) shows that the exact computation of $\nabla\mathcal{E}(\mathbb{P})$ requires to explore all subsets if size k in $\mathcal{D}$. Indeed, by looking carefully to the nature of $\mathcal{E}$ and $\nabla\mathcal{E}$, it is possible to observe that

$$\pi_\mathcal{D}\left(\nabla\mathcal{E}(\mathbb{P})\right) = \mathbb{E}_\mathbb{P}\left[\pi_\mathcal{D}\left(\frac{C(\mathcal{G}, .)\hat{q}(\hat{\mathbb{A}}_{\mathcal{G},n})}{\mathbb{P}(.)}\right)\right].$$

It is then possible to produce two stochastic algorithms which approach the behaviours of (2.4) and (2.5), and it makes possible to learn some optimal $\mathbb{P}$. Let be given some positive steps $(\alpha_j)_{j\in\mathbb{N}}$ such that

$$(\mathbf{H_0}) \qquad \sum_{j=1}^{+\infty} \alpha_j = +\infty \qquad \text{and} \qquad \exists \nu > 0 \quad \sum_{j=1}^{+\infty} \alpha_j^{1+\nu} < +\infty,$$

we can define a learning algorithm of $(\mathbb{P}_j)_{j\geqslant 0}$ (described by Algorithm (1)).

Algorithm 1 Feature selection using a stochastic gradient algorithm (approximation of (2.4)).

Require: Dictionary $\mathcal{D}$, Algorithm $\mathbb{A}$, Dataset $(X_i, Y_i)_{i\in 1\ldots n}$, integers $k \in]0; n[$ and J.
Ensure: $\mathbb{P}$ minimiser of $\mathcal{E}$
  $\mathbb{P}_0 = \mathcal{U}_{\mathcal{D}}$, uniform law on $\mathcal{D}$.
  $j \leftarrow 0$
  while $j < J$ **do**
    Sample $\mathcal{G}_j$ in $\mathcal{D}^k$ according to $\mathbb{P}_j^{\otimes k}$
    Compute $\hat{\mathbb{A}}_{\mathcal{G}_j,n}$ as well as an estimation of the classification error $\hat{q}(\hat{\mathbb{A}}_{\mathcal{G}_j,n})$
    Update the weights $\mathbb{P}_{j+1}$ as

$$\forall g \in \mathcal{D} \qquad \mathbb{P}_{j+1}(g) = \pi_{\mathcal{S}} \circ \pi_{\mathcal{D}} \left(\mathbb{P}_j - \alpha_j \left(\frac{C(\mathcal{G}_j, .)\hat{q}(\hat{\mathbb{A}}_{\mathcal{G}_j,n})}{\mathbb{P}_j} \right) \right) (g)$$

    $j \leftarrow j + 1$
  end while

If we consider now the affine time interpolation $(\mathbb{P}_t^{interp})_{t\geqslant 0}$ of $(\mathbb{P}_j)_{j\geqslant 0}$ at times

$$\tau_j = \sum_{i\leqslant j} \alpha_i,$$

it is possible to use standard results of Robbins-Monro method and show the following result (see *e.g.*[Kushner and Yin, 2003] or [Benaim, 1996]) :

Theorem 2.1.1 (Convergence of OFW (*Optimal Feature Weighting*)) *The interpolated process $(\mathbb{P}_t^{interp})_{t\geqslant 0}$ is an asymptotic pseudo-trajectory of the differential equation (2.4). Moreover, the algorithm converges to a local minimum of $\mathcal{E}$.*

It is also possible to obtain a similar result for the diffusion approximation of the stochastic algorithm (these results may be found in[5]). This more exploratory algorithm is described by Algorithme (2).

Again, if we denote $(\check{\mathbb{P}}_t^{interp})_{t\geqslant 0}$ as the continuous time affine interpolation of $(\check{\mathbb{P}}_j)_{j\geqslant 0}$ at timesτ_j, classical methods of stochastic approximation of[Kushner and Yin, 2003] or [A. Benveniste and Priouret, 198 lead to a tightness result for $(\check{\mathbb{P}}_t^{interp})_{t\geqslant 0}$ and an identification procedure shows the following result.

Theorem 2.1.2 (Convergence of the diffusive OFW) *The stochastic process $(\check{\mathbb{P}}_t^{interp})_{t\geqslant 0}$ weakly converges towards the unique invariant measure of (2.5). It is also the case for $(\check{\mathbb{P}}_j)_{j\in\mathbb{N}}$.*

The main technical difficulty of the proof relies on the underlying tightness result when a projection on the simplex $\mathcal{S}_{\mathcal{D}}$ occurs.

Figure2.2 represents as an example some subsets of features selected by using OFW on a faces dataset which is predicted by a SVM algorithm.

Algorithm 2 Feature selection using the stochastic diffusive approximation of (2.5).

Require: Dictionary $\mathcal{D}$, Algorithm $\mathbb{A}$, Dataset $(X_i, Y_i)_{i\in 1\dots n}$, integers $k \in]0; n[$ and J, variance σ^2.

Ensure: $\check{\mathbb{P}}$ minimiser of $\mathcal{E}$

$\check{\mathbb{P}}_0 = \mathcal{U}_{\mathcal{D}}$, uniform law on $\mathcal{D}$.

$j \leftarrow 0$

while $j < J$ **do**

Sample $\mathcal{G}_j$ in $\mathcal{D}^k$ according to $\check{\mathbb{P}}_j^{\otimes k}$

Compute $\hat{\mathbb{A}}_{\mathcal{G}_j,n}$ as well as an estimation of the classification error $\hat{q}(\hat{\mathbb{A}}_{\mathcal{G}_j,n})$

Sample p independent random variables $(\xi_j(g))_{g\in\mathcal{D}} \sim \mathcal{N}(0,1)^{\otimes p}$

Update the weights $\mathbb{P}_{j+1}$ as

$$\forall g \in \mathcal{D} \qquad \check{\mathbb{P}}_{j+1}(g) = \pi_{\mathcal{S}} \circ \pi_{\mathcal{D}} \left(\check{\mathbb{P}}_j - \alpha_j \left(\frac{C(\mathcal{G}_j, .)\hat{q}(\hat{\mathbb{A}}_{\mathcal{G}_j,n})}{\check{\mathbb{P}}_j} \right) + \sqrt{\alpha_j}\sigma\xi_j \right)(g)$$

$j \leftarrow j + 1$

end while

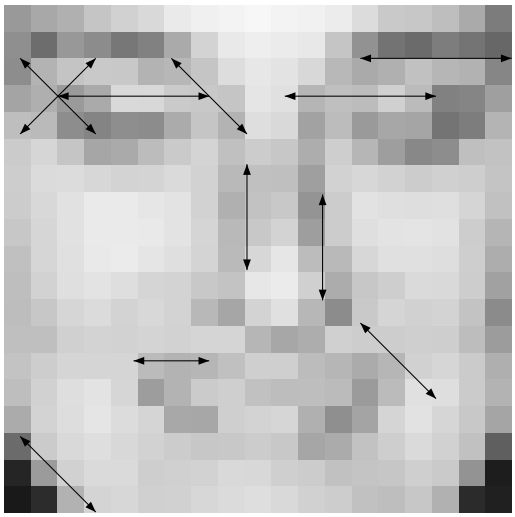


FIGURE 2.2 – Main binary edge detectors selected by OFW on a face recognition problem.

2.2 Sequential stochastic algorithm for design of experiments

In this paragraph, we describe a new stochastic method for building optimal design of experiments in order to find « good »adaptive designs for the statistical regression problem. For sake of simplicity, we will only consider the case of an unknown function η which is defined on $\Omega = [0; 1]^d$ and we aim to produce a sequential method that finds a finite number of points suitable to build a regression of η as good as possible on Ω. In the sequel, we describe our approach with $d = 1$ but this can be easily generalizes to larger dimension, as well as the associated theoretical results.

2.2.1 Framework

We assume that η belongs to an homogeneous Besov space with unknown regularity s [1]. We aim to predict η as a finite linear combination of elements taken in the dictionary $(\Lambda_{j,k})_{j\in\mathbb{N},k=0\dots 2^j-1} = \mathcal{D}$. Here, $\mathcal{D}$ is a multi-resolution analysis expanded on a wavelet decomposition and we observe only noisy version of the signal η through the computation of f

$$f(x) = \eta(x) + \sigma\xi(x), \tag{2.6}$$

where $\xi(x)$ is a normalized Gaussian noise and σ^2 is the variance of this noise which is unknown. We want to find optimal points of measurement in Ω (2.6) for which the prediction of η will be optimal.

Our method proposes to use iteratively some simple linear models computed on a small subsets of elements in $\mathcal{D}$ and the main difficulty at step n is to find an optimal design $\mathbf{x}_n$ computed with a sub-dictionary $\mathcal{D}_n$. For some rather trivial reasons linked to the framework of design of experiments, [2], we impose that

$$\mathbf{x}_{n+1} = \mathbf{x}_n \cup \{\zeta_{n+1}\}. \tag{2.7}$$

We associate to each linear model $\hat{\eta}_{\mathbf{x}_n,\mathcal{D}_n}$ a criterion which measures the quality of approximation and that corresponds to the mean integrated square error

$$J(\hat{\eta}_{\mathbf{x}_n,\mathcal{D}_n},\eta) = \int_\Omega \mathbb{E}[\hat{\eta}_{\mathbf{x}_n,\mathcal{D}_n}(u) - \eta(u)]^2 du.$$

η may be decomposed on $\mathcal{D}_n$ and its orthogonal and one can write the above criterion following a bias variance tradeoff :

$$J(\hat{\eta}_{\mathbf{x}_n,\mathcal{D}_n},\eta) = \|\mathbb{E}\hat{\eta}_{\mathbf{x}_n,\mathcal{D}_n} - \eta_{\mathcal{D}_n}\|_\Omega^2 + \|\eta_{\mathcal{D}_n^c}\|_\Omega^2 + \sigma^2 \mathrm{Tr}\left(\mu_{1,1}(\mathcal{D}_n))M_{\mathbf{x}_n,\mathcal{D}_n}^{-1}\right).$$

Event if the bias term is intractable, it is possible to compute a pessimistic estimation following a minimax approach that depends on a parameter $\tau > 0$ which quantifies the size of the bias that cannot be compressed using only elements of $\mathcal{D}_n$

$$\|\mathbb{E}\hat{\eta}_{\mathbf{x}_n,\mathcal{D}_n} - \eta_{\mathcal{D}_n}\|_\Omega^2 + \|\eta_{\mathcal{D}_n^c}\|_\Omega^2 \leqslant \sup_{\|\nu\|_{\mathcal{D}_n^c}^2 \leqslant \tau} \|\mathbb{E}\hat{\eta}_{\mathbf{x}_n,\mathcal{D}_n} - \nu_{\mathcal{D}_n}\|_\Omega^2 + \|\nu_{\mathcal{D}_n^c}\|_\Omega^2 := B^\star_{\mathbf{x}_n,\mathcal{D}_n,\tau}.$$

These simple facts yields considering the balanced minimax criterion

$$J^\star(\hat{\eta}_{\mathbf{x}_n,\mathcal{D}_n},\eta) := B^\star_{\mathbf{x}_n,\mathcal{D}_n,1} + \lambda \mathrm{Tr}\left(\mu_{1,1}(\mathcal{D}_n))M_{\mathbf{x}_n,\mathcal{D}_n}^{-1}\right), \tag{2.8}$$

where $\lambda = \sigma^2\tau^{-2}$ is a parameter which penalizes the variance of the estimation.

2.2.2 Algorithm of sequential design of experiments

In [4], we propose to use a sequential algorithm that builds $\mathbf{x}_n$ and upgrades $\mathcal{D}_n$: $\mathbf{x}_n$ aims to control the variance of estimation although $\mathcal{D}_n$ optimizes the bias of the linear model. The computation of ζ_{n+1} (see equation (2.7) relies on the optimization of $J^\star(\hat{\eta}_{\mathbf{x}_n,\mathcal{D}_n},\eta)$ and $\mathcal{D}_{n+1}$ is obtained by the addition or deletion of one sons to $\mathcal{D}_n$ following a Metropolis-Hastings strategy so that $|\mathcal{D}_{n+1}\,\Delta\,\mathcal{D}_n| = 1$. This method is described in Algorithm 3.

1. Remark that in our framework, the "adaptive" nature of the algorithm has no common point with the classical sense of adaptive estimation in mathematical statistics.

2. Each measurement of η is considered as a costly task and we do not want to throw out one measurement once it has been done

Algorithm 3 Sequential design of experiments algorithm.

Require: Dictionary $\mathcal{D}_0$, parameter $\lambda \in [0; +\infty]$, number of available measures n.
Ensure: $\mathbf{x}_n$ and $\mathcal{D}_n$
  Initialize $\mathbf{x}_0$ by minimizing (2.8).
  Compute f through measurements (2.6) and run a linear model $\hat{\eta}_{\mathbf{x}_0,\mathcal{D}_0}$.
  $j \leftarrow 0$
  while $j < n$ **do**
    Update $\mathcal{D}_{j+1}$ following the random choice
  – Addition of one son or parent of the most meaninful element of $\mathcal{D}_j$
  – Deletion of the less meaningful element in $\mathcal{D}_j$
  – Leave $\mathcal{D}_j$ unchanged
    Compute ζ_{j+1} by the minimization of (2.8).
    Measure $f(\zeta_{j+1})$ through (2.6) and upgrade the linear model $\hat{\eta}_{\mathbf{x}_{j+1},\mathcal{D}_{j+1}}$.
    $j \leftarrow j + 1$
  end while

The main idea is thus to couple a forward/backward stochastic feature selection to an adaptive choice of design of experiments. The precise description of the transition $\mathcal{D}_j \mapsto \mathcal{D}_{j+1}$ is a little bit bothersome and can be found in [4] where numerous details are given about this upgrade. Indeed, such upgrade depends on the former measure and the performance of the linear model $\hat{\eta}_{\mathbf{x}_{j+1},\mathcal{D}_{j+1}}$ at step j.

2.2.3 Results

In the former algorithme, the optimization step to compute ζ_{j+1} is the main numerical difficulty. In general, no explicit localization result is available to minimize (2.8), even in an adaptive sequential approach. In [4], we show a positive localization result which is almost explicitto compute ζ_{j+1} in the restrictive case when only the variance term is present in the criterion (2.8), thus λ is equals to $+\infty$ and for the very special case of the triangle Schauder basis. This result is described by the following theorem.

Theorem 2.2.1 *For any sub dictionary $\tilde{\mathcal{D}}$ of $\mathcal{D} = (\Lambda_{j,k})_{j=0\cdots+\infty,k=0\ldots2^j-1}$, let be given a preliminary design of experiment $\boldsymbol{x}$, then the optimal design $\boldsymbol{x} \cup \zeta$ for the criterion*

$$\mathrm{Tr}\left(\mu_{1,1}(\tilde{\mathcal{D}})M^{-1}_{\boldsymbol{x}\cup\zeta,\tilde{\mathcal{D}}}\right)$$

is obtained when ζ belongs to the set of critical points of $\tilde{\mathcal{D}}$, i.e.

$$\zeta^\star \in \cup_{\Lambda\in\tilde{\mathcal{D}}} \arg\max \Lambda.$$

This theorem is very important from a numerical point of view since it enables to build $\mathbf{x}_{j+1}$ using at the most $|\mathcal{D}_{j+1}|$ computations of the trace of the information matrix.

Moreover, when $\mathcal{D}_j$ remains fixed all along the iteration of the algorithm, it is possible to show a consistency result on the coefficient of the linear model for any multi-resolution analysis.

Theorem 2.2.2 *For any sub dictionary $\tilde{\mathcal{D}}$of $\mathcal{D} = (\Lambda_{j,k})_{j=0\cdots+\infty,k=0\ldots2^j-1}$, if $\eta = \eta_{\tilde{\mathcal{D}}} + (\eta - \eta_{\tilde{\mathcal{D}}}$ denotes the decomposition of η on $\tilde{\mathcal{D}}$, when $\lambda = +\infty$, there exists $C > 0$ such that*

$$\|\eta_{\tilde{\mathcal{D}}} - \hat{\eta}_{\boldsymbol{x}_n,\tilde{\mathcal{D}}}\| \leqslant C\sqrt{\frac{\log n}{n}}.$$

Even if the localization property (Theorem 2.2.1) concerns only a special multi-resolution analysis, it is still possible to use other basis such as Meyer wavelet basis. On a particular example, one should notice the stricking good performances of the algorithm as pointed in Figure (2.3) since only about ten points are sufficient to catch the main information in η. In [4], we also provide a numerical comparisons with penalized methods (lasso) or thresholding in wavelet basis which stand that our method behave well comparing to some other techniques.

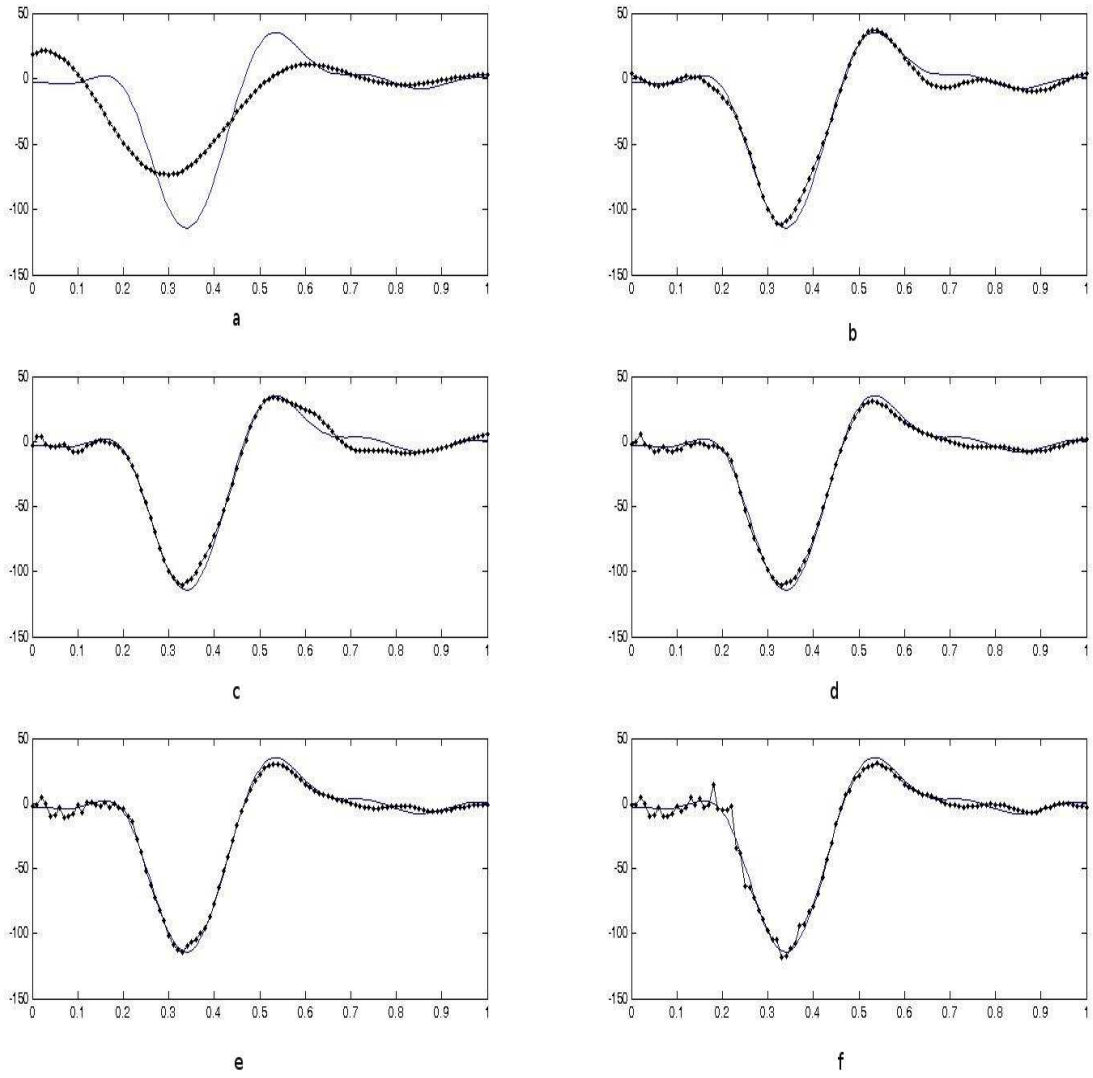


FIGURE 2.3 – Regression obtained by the sequential algorithm on the « Motorcycle » dataset with 5 points (a), 15 points (b), 25 (c), 35 (d), 45 (e), 55 (f). Continuous curve : true signal, Dashed curve : interpolation with linear model on sub dictionary of Meyer basis.

2.3 Multivariate boosting, application to gene network recovery

As pointed in the introductory paragraph, the gene network estimation can be modelled as a multivariate regression. Let be given an Hilbert space H, we aim to approach $f = (f^1, \dots f^m) \in H^{\otimes m} := H_m$ using a sequence $(G_k)_{k \geqslant 0}$. In this view, a dictionary of size p denoted $\mathcal{D}$ contains elements of H and satisfies $\overline{\text{Span}\,\mathcal{D}} = H$. In order to be consistent with real statistical applications,

the family $\mathcal{D}$ is assumed to be non orthogonal in H.

2.3.1 Brief description of Boosting algorithms

Deterministic setting $\mathbb{L}^2$-Boosting deterministic algorithms work as follows : the sequence G_k of approximation of f is initialized with $G_0 = 0$ and G_k is deduced from G_{k-1} by an improvement of prediction using a suitable unique predictor of $\mathcal{D}$. Of course, one needs to define exactly a suitable criterion to select the correct feature and the way the prediction is improved. The $\mathbb{L}^2$-Boosting method is described in Algorithm 4 for the particular case of the *Weak Greedy Algorithm* even if there exists a lot of variations around this boosting method.

Algorithm 4 Weak Greedy Algorithm (Cadre déterministe)[DeVore and Temlyakov, 1996]

Require: Dictionary $\mathcal{D}$, Function $f \in H$ to approach.

Ensure: *Shrinkage* parameter $\nu \in]0,1]$, Maximal iteration N

Predictor $G_0 = 0_H$ and Residual $R_0 = f$.

$k \leftarrow 0$

while $k < N$ **do**

Choose $\varphi_k \in \mathcal{D}$ which is sufficiently correlated with R_k $|\langle R_k, \varphi_k\rangle| \geqslant \nu \max_{g\in\mathcal{D}} |\langle R_k, g\rangle|$

Update the prediction

$$G_{k+1} = G_k + \langle R_k, \varphi_k\rangle\varphi_k$$

and the residuals

$$R_{k+1} = f - G_{k+1} = R_k - \langle R_k, \varphi_k\rangle\varphi_k$$

$k \leftarrow k + 1$

end while

Of course, the efficiency of such algorithms depends on the « size » of f. We can find in the works of [DeVore and Temlyakov, 1996] the convergence rate of G_k towards f, the size of function f is given through the constant B in the result below :

Theorem 2.3.1 ([DeVore and Temlyakov, 1996]) *Let* $B > 0$ *and assume that* $f \in \mathcal{A}(\mathcal{D}, B)$ *with*

$$\mathcal{A}(\mathcal{D}, B) = \left\{ f = \sum_{g_j\in\mathcal{D}} a_j g_j \quad \textit{such that} \quad \|a\|_1 \leqslant B \right\}.$$

There exists C_B *that only depends on* B *for which the residual* R_k *satisfies*

$$\|R_k\|_H \leqslant C_B(1 + \nu^2 k)^{-\frac{\nu}{2(2+\nu)}}.$$

The effect of size of the shrinkage parameter ν is to slow down the convergence rate which is thus optimal when $\nu = 1$ where one recovers an approximation rate of $k^{-1/6}$. Event if $\nu < 1$ seems useless in the deterministic framework, it is indeed an important feature of the algorithm for its application in a noisy setting as pointed in the next paragraph.

Random framework The approach of [Bühlmann, 2006] is to show the stability of the $\mathbb{L}^2$-Boosting in a noisy setting when one observes a n-sample $(X_1, Y_1), \dots (X_n, Y_n)$ i.i.d. where

$$\forall i \in \{1 \dots n\} \qquad Y_i = f(X_i) + \epsilon_i,$$

with the assumption that f still may be decomposed in $\overline{\text{Span}\,\mathcal{D}}$. If one denotes H the Hilbert space $\mathbb{L}^2(P)$ where P is the unknown law of the design X, we cannot access with our n sample to some empirical features of f. We define the empirical scalar product and norm as

$$\forall (h_1, h_2) \in H^2 \qquad \langle h_1, h_2 \rangle_{(n)} = \frac{1}{n} \sum_{i=1}^{n} h_1(X_i) h_2(X_i) \qquad \text{et} \qquad \|h_1\|^2_{(n)} = \langle h_1, h_1 \rangle_{(n)}.$$

The WGA may be extended to a noisy setting and is described by Algorithm5 [3].

Algorithm 5 Weak Greedy Algorithm (Noisy setting)[Bühlmann, 2006]

Require: Dictionary $\mathcal{D}$, $(X_i, Y_i)_{i \in \{1 \dots n\}}$

Ensure: *Shrinkage* parameter $\nu \in]0, 1]$, Maximal iteration N_n

Predictor $G_0 = 0_H$ and Residual $R_0 = f$.

$k \leftarrow 0$

while $k < N_n$ **do**

Choose $\varphi_k \in \mathcal{D}$ which is sufficiently correlated with the « observed »residual :

$$\left| \langle Y - G_k, \varphi_k \rangle_{(n)} \right| \geqslant \nu \max_{g \in \mathcal{D}} \left| \langle Y - G_k, g \rangle_{(n)} \right|$$

Update the prediction

$$G_{k+1} = G_k + \langle Y - G_k, \varphi_k \rangle \varphi_k$$

and the theoretical unobserved residuals

$$R_{k+1} = R_k - \langle R_k, \varphi_k \rangle_{(n)} \varphi_k - \langle \epsilon, \varphi_k \rangle_{(n)} \varphi_k.$$

$k \leftarrow k + 1$

end while

2.3.2 Boost-Boost Algorithm for multivariate regression (deterministic case)

Generalization of [Lutz and Bühlmann, 2006] In the multivariate setting, there are m coordinatesf to predict and a natural extension of the former algorithm may consider (for instance in the deterministic case) a new sequence of predictors/residuals initialized with $G_0 = 0_{H_m}, R_0 = f$ and whose iteration at step k aims to find $i_k \in \{1 \dots m\}$ and $\varphi_k \in \mathcal{D}$ such that

$$\left| \langle R_k^{i_k}, \varphi_k \rangle \right| \geqslant \nu \max_{i \in \{1 \dots m\}, g \in \mathcal{D}} \left| \langle R_k^i, g \rangle \right|.$$

Such choice of coordinate and predictor has been considered in [Lutz and Bühlmann, 2006] in a noisy setting. The main advantage of such extension is that it shortens theoretical complications concerning both deterministic and noisy cases which are thus simple adaptations of the univariate boosting. Indeed, such choice for i_k does not take into account the size of the residuals on each coordinate. This may affect the efficiency of the boosting algorithm in the noisy setting since we cannot use an infinite credit of iterations to predict each coordinate : the maximal iteration N_n theoretically depends on the size n and in practical situations one stops the iterations following AIC. It is thus important to well choose the coordinate to predict i_k to obtain an efficient algorithm.

3. Let us stress a minor mistake in [Bühlmann, 2006, Lutz and Bühlmann, 2006] where theoretical residuals $R_k = f - G_k$ (which are unobserved) are used to define the sequential predictors φ_k instead of the empirical residuals $Y - G_k$ which are only available to define the algorithm.

Boost-Boost algorithm for multivariate regressions We develop for multivariate regression a booting algorithm which spreads its effort on all the coordinates of H_m all along the iterations in order to avoid the lack described above concerning the approach of [Lutz and Bühlmann, 2006]. We propose to select the coordinate i_k with two different methods which are described in Algorithm 6.

Algorithm 6 Boost-Boost Algorithm (Deterministic case)[17]

Require: Dictionary $\mathcal{D}$, function $f \in H$ to approach.

Ensure: *Shrinkage* parameters μ, γ and ν in $]0,1]$, Maximal iteration N

Predictor $G_0 = 0_H$ and Residual $R_0 = f$.

$k \leftarrow 0$

while $k < N$ **do**

Coordinate i_k to boost

$$\|R_k(f^{i_k})\|^2 \geqslant \mu \max_{1 \leqslant i \leqslant m} \|R_k(f^i)\|^2 \qquad [\mathbb{L}^2 \text{ norm of the residuals}] \qquad (2.9)$$

or

$$\sum_{j=1}^{p} \langle R_k(f^{i_k}), g_j \rangle^2 \geqslant \mu \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{p} \langle R_k(f^i), g_j \rangle^2. \qquad [\text{Sum of correlations with } \mathcal{D}] \qquad (2.10)$$

Choice $\varphi_k \in \mathcal{D}$ sufficiently correlated with the residual R_k : $|\langle R_k, \varphi_k \rangle| \geqslant \nu \max_{g \in \mathcal{D}} |\langle R_k, g \rangle|$

Update

$$G_{k+1}^{i_k} = G_k^{i_k} + \gamma \langle R_k^{i_k}, \varphi_k \rangle \varphi_k \qquad \text{and} \qquad \forall i \neq i_k \quad G_{k+1}^i = G_k^i.$$

Update the residuals

$$R_{k+1}^{i_k} = R_k^{i_k} - \gamma \langle R_k^{i_k}, \varphi_k \rangle \varphi_k \qquad \text{and} \qquad \forall i \neq i_k \quad R_{k+1}^i = R_k^i.$$

$k \leftarrow k + 1$

end while

Hence, these algorithms proceed as follows : we first seek the best coordinate i_k (the one which is the most informative for the prediction) and then use the best predictor φ_k for this choice of coordinate. It is still possible to obain a convergence result for the boost-boost algorithm based on the $\mathbb{L}^2$ norm of the residuals as defined by (2.9). Again, this rate depends on the size of f and of the shrinkage parameters μ, γ and ν introduced in Algorithm 6.

Theorem 2.3.2 (Boost-Boost Algorithm (deterministic case and $\mathbb{L}^2$ norm of residuals)) *Let $f = (f^1, \dots f^m) \in H_m$ such that all coordinates $f^j \in \mathcal{A}(\mathcal{D}, B)$. Then, for all $k \geqslant m$, Algorithm 6 which uses (2.9) converges : there exists $C_B > 0$ which only depends on B such that*

$$\forall i \in \{1, \dots, m\}, \quad \|R_k(f^i)\| \leqslant \mu^{-\frac{1}{2}} \nu^{-\frac{\nu(2-\gamma)}{2+\nu(2-\gamma)}} (\gamma(2-\gamma))^{-\frac{-\nu(2-\gamma)}{2(2+\nu(2-\gamma))}} C_B \left(\frac{k}{m}\right)^{-\frac{\nu(2-\gamma)}{2(2+\nu(2-\gamma))}}.$$

The proof is a technical extension of the proof of [DeVore and Temlyakov, 1996], the idea is to remark with a large number of iterations, one coordinate is sufficiently chosen and this selected coordinate through (2.9) enable a global control of the residuals.

About the second boost-boost algorithm that uses the sum of correlations with $\mathcal{D}$ to find i_k (equation (2.10)), it may also be analysed with an assumption on the coherence of the dictionary

$\mathcal{D}$ defined as

$$\rho = \sup_{i \neq j, g_i \in \mathcal{D}, g_j \in \mathcal{D}} |\langle g_i, g_j \rangle|,$$

which may be related to the S sparsity of each f^j :

$$f^j = \sum_{i=1}^{p} \alpha_i^j g_i \qquad \text{with} \qquad \|\alpha_\cdot^j\|_0 \leqslant S.$$

It is then possible to obtain the following result.

Theorem 2.3.3 (Boost-Boost algorithm (deterministic case, sum of correlations with $\mathcal{D}$)) *Let $f = (f^1, \dots f^m) \in H_m$ such that each coordinate $f^j \in \mathcal{A}(\mathcal{D}, B)$ is S sparse, we assume moreover that $\rho((1+\nu^{-1})S - 1) < 1$. Then there exists $C_{\rho,S,B}$ which depends only on the size B and the coherence ρ such that for all $k \geqslant 1$,*

$$\forall i \in \{1, \dots, m\}, \quad \|R_k(f^i)\| \lessapprox \mu^{-\frac{1}{2}} \nu^{\frac{-\nu(2-\gamma)}{2+\nu(2-\gamma)}} (\gamma(2-\gamma))^{\frac{-\nu(2-\gamma)}{2(2+\nu(2-\gamma))}} C_{\rho,S,B} \left(\frac{k}{m}\right)^{-\frac{\nu(2-\gamma)}{2(2+\nu(2-\gamma))}}.$$

Some link between the coherence of $\mathcal{D}$ and the sparsity of f has already been pointed for the approximation of f using Boosting algorithms by several works ([Temlyakov and Zheltov, 2011, Tropp, 2004] for instance). More precisely, the assumption $\rho(2\delta - 1) < 1$ (obtained when $\nu = 1$ for instance) ensures that all along the iteration of the boosting algorithm, the residual R_k is at the most S sparse. In fact, the sparsity of the residual is non increasing all along the iterations of the boosting and the Boosting algorithm does not use some "wrong" elements of the dictionary. The importance of such assumption here is thus not really surprising.

2.3.3 Boost-Boost Algorithm for multivariate noisy regressions

It is still possible to adapt the former boost-boost methods described by Algorithm 6 in the realistic noisy setting. Using the notations introduced in paragraph 2.3.1 concerning empirical data, the boost-boost method is developed by Algorithm7.

One should remark that for the $\mathbb{L}^2$ norm of residuals, no shrinkage can be used for μat least from a theoretical point of view. It is still possible to show the statistical consistency following the ideas given in [Bühlmann, 2006]. The main idea is to consider a « phantom »algorithm which would work in a deterministic setting with the several selections made by its stochastic version. Note that indeed, one may consider a dictionary $\mathcal{D}$ which is composed of a set of variables p_n which may increase with the number of samples n in the dataset. Several assumption are needed to obtain statistical consistency.

The first assumption is a technical hypothesis both on the structure of the design X and the dictionary $\mathcal{D}$. This technical condition is necessary to obtain uniform Law of Large Number results.

Assumption 1 ($\mathbf{H}_{\mathcal{D}}$) *For any choice of predictor g in the dictionary $\mathcal{D}$, the random variable $g(X)$ has a normalized second order moment and is essentially bounded*

$$\forall j \in \{1 \dots p_n\} \qquad \mathbb{E}[g_j(X)^2] = 1 \quad \textit{and} \quad \sup_{1 \leqslant j \leqslant p_n, n \in \mathbb{N}} \|g_j(X)\|_\infty < \infty.$$

The next hypothesis defines the exact very large dimension setting where it is possible to estimate something in the model as soon as $\log p << n$.

Algorithm 7 Boost-Boost algorithm (random case)[17]

Require: Dictionary $\mathcal{D}$, function $f \in H$ to estimate.

Ensure: *Shrinkage* parameters μ, γ and ν in $(0,1]$, Maximal iteration N_n

Predictor $\hat{G}_0 = 0_H$ and Residual $R_0 = f$.

$k \leftarrow 0$

while $k < N$ **do**

Choice of the coordinate i_k to boost

$$\|Y - \hat{G}_k^{i_k}\|_{(n)}^2 \geqslant \max_{1 \leqslant i \leqslant m} \|Y - \hat{G}_k^i\|_{(n)}^2 \qquad [\ \mathbb{L}^2 \text{ norm of the residuals}] \qquad (2.11)$$

or

$$\sum_{j=1}^{p} \langle Y - \hat{G}_k^{i_k}, g_j \rangle_{(n)}^2 \geqslant \mu \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{p} \langle Y - \hat{G}_k^{i_k}, g_j \rangle_{(n)}^2. \qquad [\text{Sum of correlations with } \mathcal{D}] \qquad (2.12)$$

Choice of $\varphi_k \in \mathcal{D}$ sufficiently correlated with the empirical residual $Y^{i_k} - \hat{G}_k^{i_k}$:

$$\left| \langle Y^{i_k} - \hat{G}_k^{i_k}, \varphi_k \rangle \right|_{(n)} \geqslant \nu \max_{g \in \mathcal{D}} \left| \langle Y^{i_k} - \hat{G}_k^{i_k}, g \rangle \right|_{(n)}$$

Update the predictor

$$\hat{G}_{k+1}^{i_k} = \hat{G}_k^{i_k} + \gamma \langle R_k^{i_k}, \varphi_k \rangle \varphi_k \qquad \text{et} \qquad \forall i \neq i_k \quad \hat{G}_{k+1}^i = \hat{G}_k^i.$$

Update the unobserved residuals

$$R_{k+1}^{i_k} = R_k^{i_k} - \gamma \langle R_k^{i_k}, \varphi_k \rangle_{(n)} \varphi_k - \langle \epsilon^{i_k}, \varphi_k \rangle_{(n)} \varphi_k \qquad \text{et} \qquad \forall i \neq i_k \quad R_{k+1}^i = R_k^i.$$

$k \leftarrow k + 1$

end while

Assumption 2 ($\mathbf{H_{p_n}}$) *The number of regressors p_n in $\mathcal{D}$ satisfies*

$$p_n = \underset{n \to +\infty}{O} \left(\exp(C n^{1-\xi}) \right),$$

for some $\xi \in]0,1[$ and a constant $0 < C < \infty$.

The main assumption comes from a sparse structure of the signal f to recover. It is described by the next hypothesis and is obviously true as soon as the sparsity index remains fixed when n is growing to $+\infty$.

Assumption 3 ($\mathbf{H_f}$) *The function $f = (f^1, \ldots, f^m)$ to predict is spanned in H_m*

$$\forall j \in \{1 \ldots m\} \qquad f^j = \sum_{i=1}^{p_n} \gamma_i^{(j)} g_j$$

and each coordinate f^j is S sparse with S independent from n, which implies that the sequence $(\gamma_i^{(j)})_{n \in \mathbb{N}, 1 \leqslant j \leqslant m, 1 \leqslant i \leqslant p_n}$ satisfies

$$\forall 1 \leqslant j \leqslant m \qquad \sup_{n \in \mathbb{N}} \sum_{i=1}^{p_n} |\gamma_i^{(j)}| < \infty.$$

At last, the next hypothesis is on the noise structure : we must have a sufficiently large order bounded moment to use thresholding argument coupled with Bernstein's inequality to sharply control the difference between $\langle , \rangle$ and $\langle , \rangle_{(n)}$.

Assumption 4 ($\mathbf{H}_\epsilon$) *The random variables which model the noise $(\varepsilon_\ell)_{\ell=1\dots n}$ are i.i.d. centered in $\mathbb{R}^m$ and of covariance Id_m, independent on the $(X_\ell)_{\ell=1\dots n}$ such that*

$$\sup_{1\leqslant j\leqslant m, n\in\mathbb{N}} \mathbb{E}|\varepsilon^{(j)}|^s < \infty,$$

for any $s > \frac{2}{\xi}$ where ξ is given in assumption 2.

This assumption is satisfied as soon as the tail of the noise distribution is of Gaussian or Laplace nature.

At last, the next assumption is on the magnitude of the active coefficients in view to obtain a consistency result for the support recovery problem.

Assumption 5 ($\mathbf{H}_\mathcal{S}$) *The active coefficients in the S sparse representation of each coordinate f^i satisfy :*

$$|\gamma_i^{(j)}| \geqslant n^{-\kappa\xi},$$

with $\kappa < 1/2$.

We now state the support recovery result of the Boost-Boost algorithms.

Theorem 2.3.4 (Boost-Boost support recovery) *The next three points are satisfied with large probability :*

i) *Suppose that assumptions 1-4 are fulfilled ($(\mathbf{H}_\mathcal{D}), (\mathbf{H}_{\mathbf{p_n}}), (\mathbf{H}_\mathbf{f}), (\mathbf{H}_\epsilon)$) ,and that each coordinate f^j is S sparse with $\rho((1+\nu^{-1})S-1) < 1$. Then there exists a maximal explicit value γ^* of the shrinkage parameter γ and a growing number of stopping iterations $(k_n)_{n\in\mathbb{N}}$ such that for any $0 < \gamma < \gamma^*$, the Boost-Boost algorithm based on the $\mathbb{L}^2$ norm of residuals only selects "good" coefficients.*

ii) *Assume moreover that the hypothesis (5) $\mathbf{H}_\mathcal{S}$ holds, then there exists a maximal value $\kappa^*(\gamma, S)$ such that if $\kappa \leqslant \kappa^*(\gamma, S)$, Boost-Boost algorithm recovers the support of f.*

iii) *If one supposes the strongest hypothesis that $p_n = \underset{n\to+\infty}{o}(\sqrt{n})$, the result still holds for the Boost-Boost algorithm based on the sum of correlations with $\mathcal{D}$ (for a different value of γ^*).*

This last result were known for other kind of sparse reconstruction algorithm. The threshold $n^{-\xi/2}$ corresponds to the minimal value of the amplitude of active variables. Below this threshold, it seems impossible to consistently estimate the support of f. When such hypothesis is not satisfied, it is however to show that only "good" variables are built by the Boost-Boost algorithm. This point is not true in general for other sparse algorithms and is provided here by the shrinking parameter γ which allows to obtain thus a slightly stronger result. Note that for the Boost-Boost algorithm based on the $\mathbb{L}^2$ norm of residuals, $\gamma^* \simeq 13/18$ for instance, and this value is lower for the other Boost-Boost algorithm owing to poorer concentration properties of the sum of correlations with $\mathcal{D}$ of residuals.

Such result permits to obtain the next theorem, and we should note that the hypothesis (5) $\mathbf{H}_\mathcal{S}$ is not yet necessary to obtain such consistencies.

Theorem 2.3.5 (Boost-Boost consistency) *Suppose that assumptions 1-4 are fulfilled ($(\mathbf{H}_{\mathcal{D}}), (\mathbf{H}_{\mathbf{p_n}}), (\mathbf{H}_{\mathbf{f}}), (\mathbf{H}_{\epsilon})$) ,and that each coordinate f^j is S sparse with $\rho((1+\nu^{-1})S-1) < 1$, then there exists a sufficiently slow increasing number of iteration $(k_n)_{n\in\mathbb{N}}$ whose limit is $+\infty$ such that the Boost-Boost algorithm based on the $\mathbb{L}^2$ norm of residuals satisfies*

$$\forall i \in \{1,\dots,m_n\}, \quad \mathbb{E}\|f - \hat{G}_{k_n}(f)\|^2 = \underset{n\to+\infty}{o_P}(1)$$

as soon as $\gamma < \gamma^*$. If one suppose the strongest hypothesis that $p_n = \underset{n\to+\infty}{o}(\sqrt{n})$, the result still holds for the Boost-Boost algorithm based on the sum of correlations with $\mathcal{D}$:

$$\forall i \in \{1,\dots,m_n\}, \quad \mathbb{E}\|f - \hat{G}_{k_n}(f)\|^2 = \underset{n\to+\infty}{o_P}(1).$$

We should remark that the number of variables may growth exponentially fast with the number of samples thanks to a uniform law of large numbers and the assumption $(\mathbf{H}_{\epsilon})$. However, note also that the number of iterations k_n is of a logarithmic order in n. Hence, this result (which is comparable to the one of [Bühlmann, 2006]) is quite weak comparing to other results on penalized regressions such as the Lasso. We refer to[17] for some technical details on the proof of these last theorems.

2.3.4 Numerical results

We briefly describe in this paragraph numerical results obtained *via* boosting algorithms and refer to [17] or [18] for further details.

The first simulation study concerns a toy dataset already used in [Lutz and Bühlmann, 2006]. We observe a response matrix Y of size $n \times m$ and features are described in X which is a $n \times p$ matrix. The model used to generate data is $Y = X\theta + \epsilon$ where ϵ is a Gaussian noise $\mathcal{N}(0, I_n)$. Moreover, some correlations are introduced between each pair of variables (g_j, g_k) (when $1 \leqslant j, k \leqslant p$) so that $\rho(g_j, g_k) = 0.9^{|k-j|}$. Each column of θ will be s-sparse.

The second simulation studies more precisely the case of gene network inference of size $p \times p$ (p genes in the network). Each expression level of the genes is given for the n observations $(E_i)_{1\leqslant i\leqslant n} \in \mathbb{R}^{p\times n}$. The network is assumed to be self-regulated so that the following model $E = E\theta + \epsilon$ with θ the unknown matrix of regulation we aim to recover. Of course, we impose that the diagonal of θ is null to avoid trivial regression. We also assume that ϵ is a Gaussian noise $\mathcal{N}(0, I_n)$.

The last simulation is similar to the first one but we use different sparsity index for each coordinate of θ. Moreover, we introduce significantly larger correlations (± 0.9) than the ones which are present on the first data.

The precedent Figures show the evolution of the accuracy of the algorithms with respect to their power of recovery. Thus, on the abscissa, one can see the rate of coefficients which are recovered by the methods as well as the ordinate is showing the rate of good predictions. Thus, they present performances on the support recovery and do not provide any conclusion on the $\mathbb{L}^2$ error performed by the regressions. One should find in [17] complementary results.

The two first datasets show that in current situations, all boosting algorithms behave in a similar way and their numerical performances is at the least as good as the ones of classical methods such as Random Forest or Bootstrap Lasso. Moreover, our numerical studies let us think that Bayesian networks inference are a little bit less efficient than the methods used above (this is not shown in the last Figures but may be find in [17] or [18]). At last, one should remark that the first toy dataset is not a large dimensional one since the size of the feature space is 40 and we observe 50 samples.

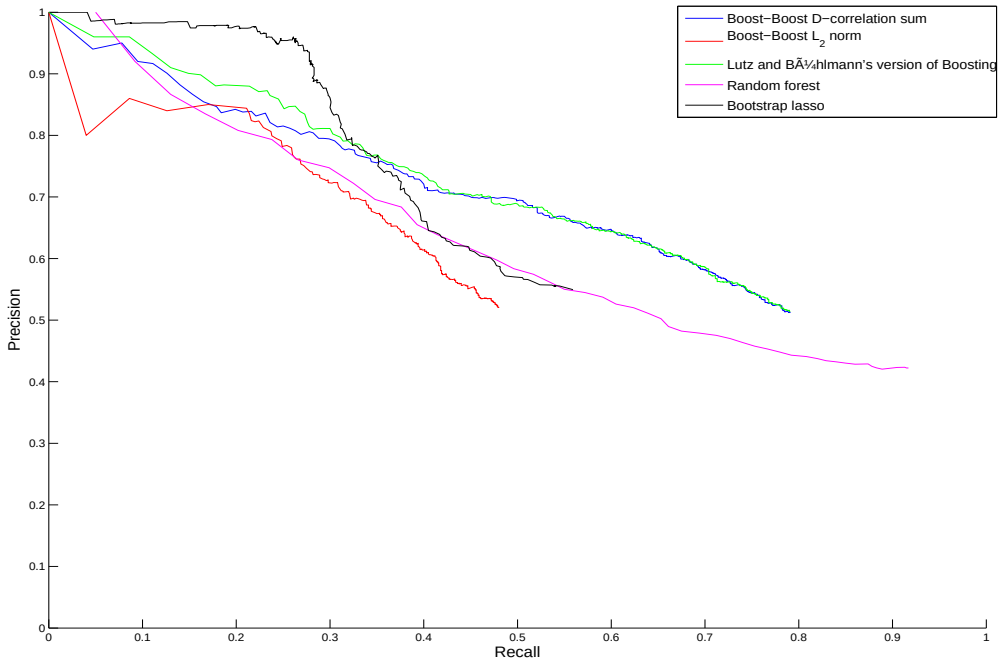


FIGURE 2.4 – Results of 5 methods of sparse regression on the toy dataset of [Lutz and Bühlmann, 2006]. We set $p = 10$, $n = 50$, $m = 4$ and the sparsity index s equals to 2. On the abscissa : the recall of reconstruction, in ordinate : the precision.

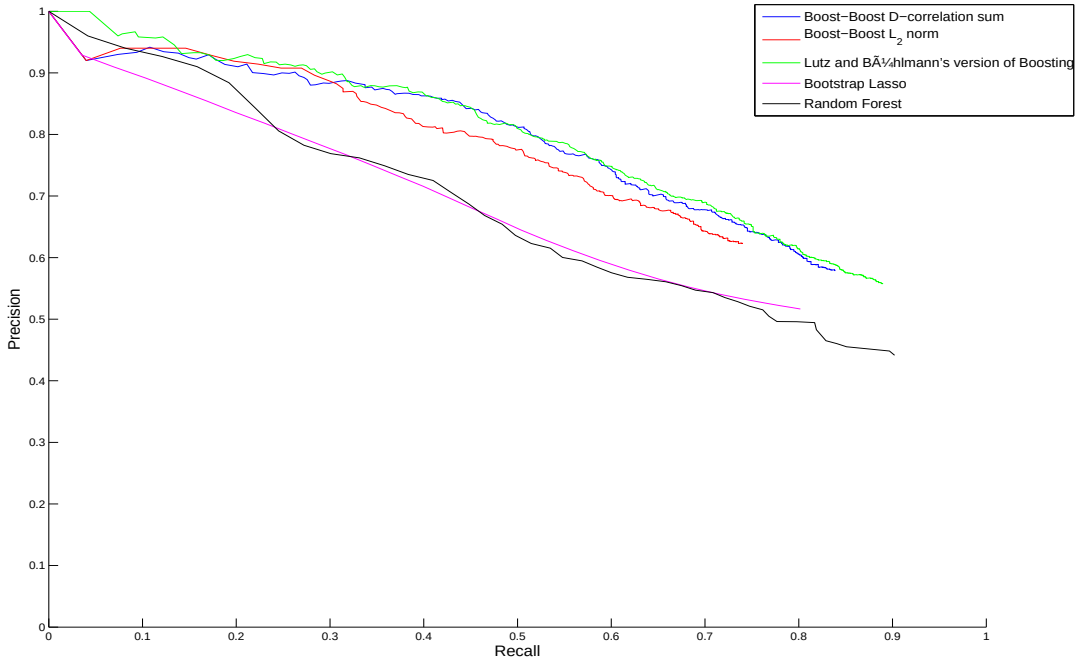


FIGURE 2.5 – Results of 5 methods of sparse regression for the gene network inference. We set $p = 10$ and $n = 50$. On the abscissa : the recall of reconstruction, in ordinate : the precision.

In the more extreme case where correlations are $\pm\frac{9}{10}$ in the last third dataset (Figure 2.6), we remark that the boost-boost sum of correlations is quite more efficient than the two other boos-

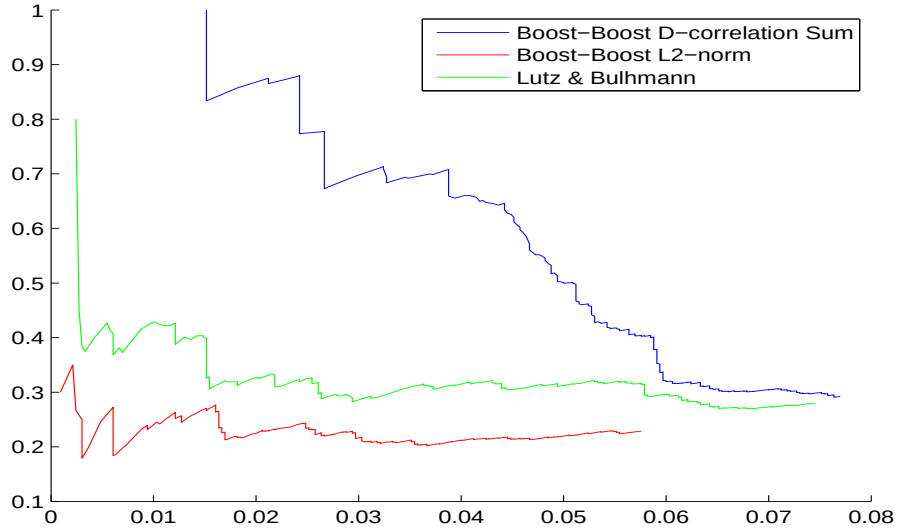


FIGURE 2.6 – Performances of the 3 boosting methods on the third model of regression with highly correlated variables with $m = 4$. Here $p = 250$, $n = 50$ as well as $m = 4$ and $s = (30, 100, 100, 100)$. On the abscissa : the recall of reconstruction, in ordinate : the precision.

ting algorithms and spreads the stress of the first iterations well comparing to other methods. Moreover, we should note that to keep a satisfactory accuracy, the algorithm only obtain a recall of $5/100$ (!) but this kind of result is obtained in the framework of the very high dimension since $n = 50$ although the size of the feature space is of the order of thousand.

2.3.5 Future works

Even though theoretical results seem satisfactory, the numerical abilities of the boosting methods for very high dimensional setting are a little bit disapointing as soon as the sparsity index S is not so small (this phenomenon is illustrated by Figure 2.6). This is slightly an expected feature of such methods when we consider the theoretical results of Theorem 2.3.4 since we should have a balance between S, p, n, γ and κ. It would be quite fair to use also some Bayesian inference approaches to deal with such problems. One should consider for instance the recent works of [Castillo and van der Vaart A., 2012] which provides some interesting enlightenments on how to use such approaches for regression in large dimensional setting.

Chapitre 3

Statistical deformable models and signal processing

In this chapter, I will provide a structured summary of the problems I studied and the solutions I found in the field of deformable statistical models. We aim to propose new methods of estimation in signal and image processing studied in a functional framework. Hence, the unknown objects to estimate belong to an abstract space $\mathcal{H}$ which will become more precise latter.

3.1 Deformation model

These problems are all concerned by observations which are corrupted by twice source of noise : the first one is a random deformation of a « mean »shape $f^\star$ and the second noise is an additive measurement noise. Each one of the n observations is described through

$$\forall i = 1 \dots n \qquad Y_i = f_i + \varepsilon_i, \tag{3.1}$$

where ε_i is the additive measurement noise, f_i is the randomly warped shape which belongs to $\mathcal{H}$ and Y_i are the final noisy observation. If $\mathcal{H}$ is a set of maps defined on Ω, each f_i are defined following the ideas of [Grenander, 1993a] by the following equation :

$$\forall x \in \Omega, \forall i = 1 \dots n \qquad f_i(x) = (f^\star + Z_i)[g_i.x]. \tag{3.2}$$

Here Z_i is a photometric variation and g_i is the action of deformation on Ω. If Z_i acts in a linear way, it is not the case for the action of g_i which is an injection of Ω into Ω. In the sequel, I will restrict the study to the case $Z_i = 0$ since no amplitude variations have been considered in my works.

3.1.1 Rigid deformation

Roughly speaking, one can dissociate two different classes of homeomorphic deformations of Ω, rigid ones and elastic ones. Rigid deformations are the simplest ones and correspond to a finite dimensional Lie group which acts on Ω. One typical example is the case where $\Omega = \mathbb{R}^d$ where we consider a group of translations. In this simple case, observations are then given by

$$\forall x \in \Omega, \forall i = 1 \dots n \qquad f_i(x) = f^\star(x - \tau_i),$$

where τ_i are random parameters of each translation to obtain f_i. The action of G can be summarized as $g.x = x - g$ for all x in Ω.

Of course, such situation may generalizes to more complex models of deformation of Ω when the group G has a larger dimension to obtain both rotations, translations, homothéties, ...

3.1.2 Elastic deformation

The second class of bijective transformation is clearly much more complex and enables to define some « elastic »deformations of Ω. These models are introduced *via* flows of differential equations by[Miller and Younes, 2001, Trouvé and Younes, 2005].

In order to model such bijective deformations of Ω, the idea is as follows : let v_i a continuous map $\mathcal{C}(\Omega, \Omega)$ and ϵ a small non negative real, the application $\phi_1 = Id + \epsilon v_i$ is always an homeomorphism as well as $\phi_p \circ \phi_{p-1} \circ \cdots \circ \phi_1$ which also reminds bijective. Finally, if one remarks that

$$\frac{\phi_p - \phi_{p-1}}{\epsilon} = v_p(\phi_{p-1}),$$

the natural generalization of small deformations $Id + \epsilon v$ depends on a family $(v_t)_{t\in[0;1]}$ of continuous maps in $\mathcal{C}(\Omega, \Omega)$ used to consider

$$\forall t \in [0;1] \qquad \frac{d\phi_t}{dt} = v_t(\phi_t)/ \tag{3.3}$$

Indeed, (3.3) admits a unique solution $(\phi_t)_{t\in[0;1]}$ given an initialization $\phi_0 = Id_\Omega$ as soon as $\int_0^1 \|v_s\| ds < +\infty$. Moreover, for all time t, ϕ_t is an homeomorphism from Ω to $\phi_t(\Omega)$. In order to keep the surjectivity of such deformations of Ω, it is enough to impose ϕ_t to be the identity on $\partial\Omega$. This last point is true as soon as $\forall t \in [0;1], \forall x \in \partial\Omega v_t(x) = 0$.

Hence, we have in our hands two very different ways to model randomly warped observations f_i from an initial mean pattern $f^\star$. Of course, in some practical cases, the second model of elastic deformation is more appropriate than the first one, and of course much more theoretical difficult to study.

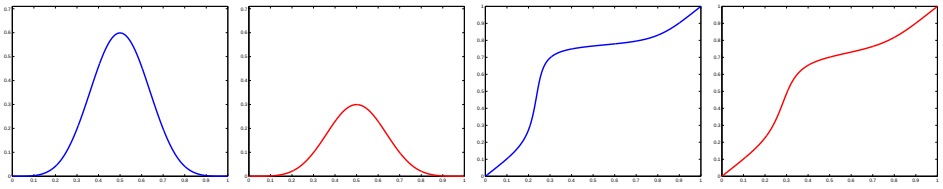


FIGURE 3.1 – Example of one dimensional homeomorphims ϕ_1 of $[0;1]$ (on the right) generated by time homogeneous vector fields v (on the left).

3.1.3 Isotonic (constrained) regression

My first statistical work on this fields concerns a simple remark that occurs in dimension 1. Indeed, a diffeomorphism of $\mathbb{R}$ is necessarily monotone. Thus, we use this last point to parametrise any monotone functions as the solution at time 1 of an O.D.E. and plug this parametrisation in a regression setting. This yields in [10] a new way to handle monotone regession.

The link between monotone functions and diffeomorphisms generated through O.D.E. is detailed below. If one denotes I any interval of $\mathbb{R}$, we define $\mathcal{H}^m(I)$ the Sobolev space

$$\mathcal{H}^m(I) = \{f : I \to \mathbb{R}, f^{(m-1)} \text{ is continuous on I and } \int_I |f^{(m)}(x)|^2 dx < +\infty\},$$

and one has the following parametrisation of monotone functions for $I = [0, 1]$ (for instance).

Theorem 3.1.1 *If* $\tilde{\mathcal{H}} = \mathrm{Span}\{1, x\} + \mathcal{H}^m(\mathbb{R})$ *and* $m \geqslant 2$. *For all non decreasing* $f \in \mathcal{H}^m([0,1])$, *define the trajectory*

$$\phi_t(x) = tf(x) + (1-t)x, \forall t \in [0,1].$$

Then, there exists a vector field $(v_t^f)_{t\in[0,1]}$ *such that* $v_t^f \in \tilde{\mathcal{H}}, \forall t \in [0;1]$ *and*

$$f = \phi_1 = \phi_0 + \int_0^1 v_t^f(\phi_t)dt.$$

Moreover, for all $t \in [0,1]$,*one has*

$$v_t^f(\phi_t(x)) = f(x) - x \text{ for all } x \in [0,1]. \tag{3.4}$$

Our idea is now to compute an estimation of f (which corresponds to ϕ_1) by an estimation of $(v_t)_{t\geqslant 0}$, and the use of (3.3) in order to obtain a naturally monotone estimate of ϕ_1. For all $t \in [0;1]$, Theorem 3.1.1 shows that v_t « maps » $tf(x) + (1-t)x$ to $f(x) - x$.

From a statistical point of view, we observe a sample of n datas $(x_1, y_1), \dots (x_n, y_n)$ such that

$$y_i = f(x_i) + \epsilon_i,$$

and $(\epsilon_i)_{i\in\{1\dots n\}}$ are centered random variables of variance σ^2. We look for a monotone $\hat{f}_n$ such that its quadratic risk defined by

$$R(\hat{f}_n, f) = \frac{1}{n}\sum_{i=1}^{n}[\hat{f}_n(x_i) - f(x_i)]^2,$$

is weak. Our strategy is to use a plug-in trick : we first compute an unconstrained estimator $\hat{f}_n^0$ of f, and then we « monotonise » to obtain $\hat{f}_n^c$ which inherits of the same theoretical asymptotic properties of $\hat{f}_n^0$. This step replaces for us the « projection »step of classical works on isotonic regression. In this view, we compute an estimation $v^{n,\lambda} = (v_t^{n,\lambda})_{t\in I}$ of $(v_t)_{t\geqslant 0}$ such that $t \in [0;1]$, $v_t^{n,\lambda}$ belongs to $\tilde{\mathcal{H}}$:

$$\forall x \in I \qquad v_t^{n,\lambda}(x) = a_1^t + a_2^t x + h_t(x), \qquad \text{where} \qquad h_t \in \mathcal{H}.$$

If we consider on $\tilde{\mathcal{H}}$ a Reproducing Kernel Hilbert Space structure described through a kernel K, a suitable way to find $v_t^{n,\lambda}$ is to solve the optimization problem

$$v_t^{n,\lambda} = \arg\min_{v\in\tilde{\mathcal{H}}} \frac{1}{n}\sum_{i=1}^{n}[(\hat{f}_n^0(x_i) - x_i) - v(t\hat{f}_n^0(x_i) + (1-t)x_i)]^2 + \lambda\|h_t\|_K^2. \tag{3.5}$$

It is then possible (see Theorem 5.1 in [10]) to obtain under some technical assumptions on K and the penalization coefficient λ_n that with large probability :

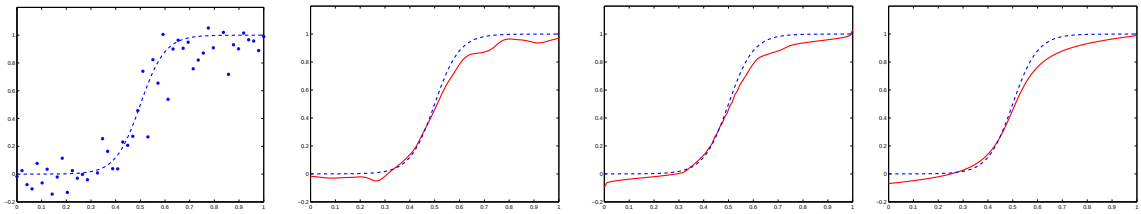
$$R(\hat{f}_n^c, f) \leqslant C(R(\hat{f}_n, f) + \lambda_n).$$

This result may also be extended to the quadratic risk on $[0;1]$.

Theorem 3.1.2 *If* $f \in \mathcal{H}^m(I)$ *satisfies* $f' > 0$ *on* I *and if one chooses* $\lambda_n = 1/n$, *then* $\hat{f}_n^c$ *built from the unconstrained* $\hat{f}_n^0$ *introduced in[Speckman, 1985], is monotone and asymptotically optimal in the minimax sense :*

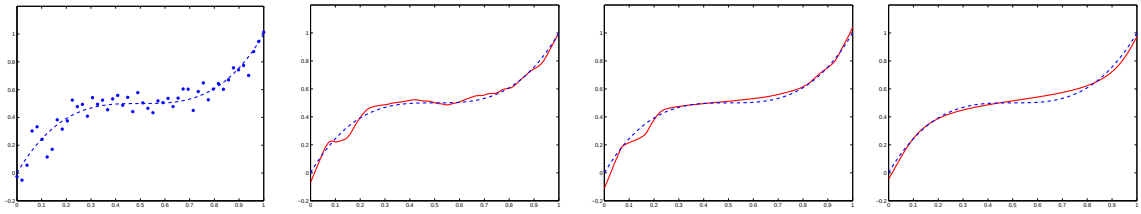
$$R_n(\hat{f}_n^c, f) = O(n^{-2m/(2m+1)}).$$

Figures (3.2), (3.3) and (3.4) are used in the experimental study presented in [10] and they all show the efficiency of the monotonisation of the unconstrained estimator through the differential flow of vector fields. One should also remark that the works presented in [10] may also be extended to larger dimensions for landmarks matching problems.



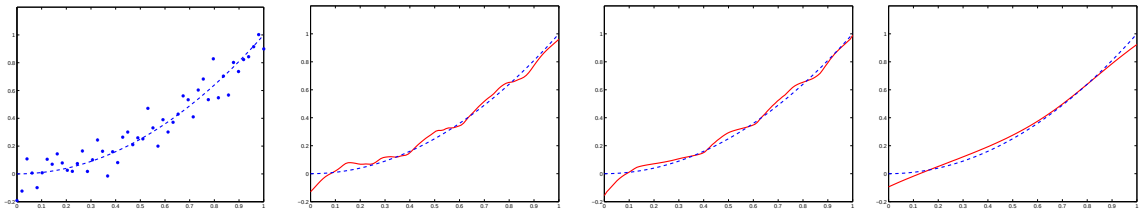
(a) (b) (c) (d)

FIGURE 3.2 – Signal m_1 : Dashed line unknown f, (a) Dataset with SNR = 3, (b) Unconstrained estimator $\hat{f}_n^0$, (c) Dette et al. estimator, (d) Monotonised estimator $\hat{f}_n^c$ from $\hat{f}_n^0$.



(a) (b) (c) (d)

FIGURE 3.3 – Signal m_2 : Dashed line unknown f, (a) Dataset with SNR = 3, (b) Unconstrained estimator $\hat{f}_n^0$, (c) Dette et al. estimator, (d) Monotonised estimator $\hat{f}_n^c$ from $\hat{f}_n^0$.



(a) (b) (c) (d)

FIGURE 3.4 – Signal m_3 : Dashed line unknown f, (a) Dataset with SNR = 3, (b) Unconstrained estimator $\hat{f}_n^0$, (c) Dette et al. estimator, (d) Monotonised estimator $\hat{f}_n^c$ from $\hat{f}_n^0$.

3.2 Deformable model with known deformation law

In this section, I provide some details on the estimation of f in the random deformation model described by equation (3.1) when one uses a white noise Gaussian model and each f_i are defined by (3.2). We aim to precisely describes what happens when the number of samples n growths to $+\infty$. We assume in this section the deformation law g known. This assumption is the most important one of the paragraphs below.

3.2.1 Randomly shifted curves

This model is of the course much most simple from a technical point of view compared to its generalizations (more complex deformation, poissonian noise) but already presents the main ideas of the estimation considered also in[8] and [16].

We assume f_i to simply be a realization of a random translation applied to the unknown f which is assumed 1-periodic. Hence, we consider the model

$$\forall j \in \{1 \dots n\} \quad \forall x \in [0;1] \qquad dY_j(x) = f(x-\tau_j)dx + \epsilon dW_j(x) \quad \text{where} (\tau_j)_{j\in\{1\dots n\}} \text{i.i.d.} \sim g. \tag{3.6}$$

Deconvolution Approach Model (3.6) may be studied by considering a Fourier basis $(e_k)_{k\in\mathbb{Z}}$ on which each Brownian motion $W_i(x)$ may be decomposed using independent gaussian coefficients. More precisely

$$\forall j \in \{1 \dots n\} \quad \forall k \in \mathbb{Z} \qquad \theta_{j,k} := \langle Y_j, e_k \rangle = \langle f, e_k \rangle e^{-i2\pi k\tau_j} + \epsilon\epsilon_{j,k},$$

where $(\epsilon_{j,k})_{j,k}$ are i.i.d. $\mathcal{N}(0,1)$.

The method to build an estimation of f is now clear when one knows Fourier coefficients $(\gamma_k)_{k\in\mathbb{Z}}$ of g : we can approach Fourier coefficients of f following the simple remark

$$\forall k \in \mathbb{Z} \qquad c_k(f) := \langle f, e_k \rangle = \frac{\langle f \star g, e_k \rangle}{\langle g, e_k \rangle} \simeq \frac{\frac{1}{n}\sum_{j=1}^{n} \theta_{j,k}}{\gamma_k}. \tag{3.7}$$

Such equality is true as soon as $\gamma_k \neq 0$ and the Strong law of Large Number should permit to well approach f. Indeed, the inversion of γ_k may become dangerous when k is large since Riemann-Lebesgue Lemma would guaranties that $\gamma_k \mapsto 0$ when $k \to \pm\infty$ and g is regular. We face here a classical phenomenon encountered in ill-posed inverse problems, which is rather logical owing to our estimation method using a deconvolution approach : we aim to invert the convolution operator for which γ_k are eigenvalues.

We should also remark that this inverse problem framework might not be the natural way to study (3.6) since we artificially expanded the problem in a Fourier analysis to obtain (3.7). We will show in the sequel that actually, the nature of estimation (3.6) as an inverse problem cannot be avoided.

Thresholding estimation and reconstruction rates on Besov space In this model, we compute the convergence rate for the mean quadratic risk : for any estimator $\hat{f}$, we define this risk as

$$\mathcal{R}(\hat{f}, f) = \mathbb{E}\|\hat{f} - f\|_2^2.$$

It is possible to build an estimator $\hat{f}_n$ from a multi-resolution analysis. More precisely, if $(\psi_{j,k})_{j,k}$ and $(\phi_{j,k})_{j,k}$ are the scaling and mother functions of Meyer wavelet decomposition at scale j and location k, we will build $\hat{f}_n$ as

$$\hat{f}_n = \sum_{k=0}^{2^{j_0}-1} \hat{c}_{j_0,k}\phi_{j_0,k} + \sum_{j\geqslant j_0}\sum_{j=0}^{2^j-1} \hat{\beta}_{j,k}\psi_{j,k},$$

where $\hat{c}$ and $\hat{\beta}$ must be estimate from the observations. The complete description of such estimation is rather technical and we will omit the details in this manuscript (they can be found in[9]). The main idea is to limit the size of j_0 and j_1 which depends on n and ν the regularity index of g. Then, we keep only coefficients whose size is greater than a threshold which is data-driven.

We prove in [9] a convergence rate of such thresholding estimator. In a simplified version presented here, this theorem is as follows.

Theorem 3.2.1 *Let $f \in B^s_{2,2}$ unknown of regularity index s unknown too, and assume that the known Fourier coefficients of g satisfy*

$$\exists\, (C_{min}, C_{max}) \quad \forall k \in \mathbb{Z} \qquad C_{min}|\ell|^{-\nu} \leqslant |\gamma_\ell| \leqslant C_{max}|\ell|^{-\nu}. \tag{3.8}$$

The thresholding estimator $\hat{f}^H_n$ described in[9] is consistent and

$$\sup_{f \in B^s_{2,2}} \mathcal{R}(\hat{f}^H_n, f) = \mathcal{O}\left(n^{\frac{-2s}{2s+2\nu+1}} \log^{\frac{2s}{2s+2\nu+1}}\right).$$

The former estimator is based on a Hard Thresholding technique and is *adaptive* to the unknown regularity s of f, this adaptivity is obtained by the use of wavelet decomposition. The main property of such multi-resolution analysis is that Meyer wavelets are band limited in Fourier analysis. Thus, our estimator is obtained through a preliminary estimation of all Fourier coefficients of f which are next plugged into the Meyer basis, we then obtain a best approximation of f.

At last, we also remark that the convergence rate $n^{\frac{-2s}{2s+2\nu+1}}$ obtained in Theorem3.2.1 is classical in statistical deconvolution regression when the curves are the realisation of $f \star g$ corrupted by a white noise :

$$\forall j \in \{1 \dots n\} \quad \forall x \in [0;1] \qquad dY_j(x) = f \star g(x)dx + \epsilon dW_j(x).$$

It is quite natural to obtain similar convergence rates since we use the same inversion of convolution operator in (3.7) as a preliminary estimation which is plugged in the computation of the Meyer wavelet coefficients.

Minimax rate of convergence In non parametric statistics, a standard method to measure the efficiency of an estimation method is to compute a lower bound of estimation and one expects that this lower bound matches asymptotically the upper bound reached by the estimator. This lower bound is described through the Minimax rate of convergence when one uses n observations to compute an estimation in a class of function $\mathcal{F}$. We then define

$$\mathcal{R}_n(\mathcal{F}) = \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} R(\hat{f}_n, f),$$

where $\hat{f}_n$ explores all possible measurable functions of the data. Hence, $\mathcal{R}_n(\mathcal{F})$ represents the best achievable rate for the worst function to estimate in the class $\mathcal{F}$.

Usually, the computation of lower bounds is much harder than the study of the upper bound of the risk with a suitable estimator. One is attempted to study a Likelihood ratio between two hypothesis which must be far enough from a metric point of view (with L^2 norm for instance) and rather near from a statistical point of view. One should refer to three classical methods to obtain such results. The most popular one is certainly the use of Fano's Lemma (introduced in [Ibragimov and Has′minskiĭ, 1981]), but sometimes other methods such as Assouad's Lemma or Le Cam's method are more tractable. In [9], we tackle the problem of the lower bound for $\mathcal{R}_n(\mathcal{F})$ when $\mathcal{F}$ is Besov space$B^s_{2,2}$ using Assouad's Lemma which is written below. One should refer to [Tsybakov, 2003] for a large description of several variations around this lemma.

Lemma 3.2.1 (Assouad's Lemma ([Bretagnolle and Huber, 1979])) *Let be given a set of functions $(f_\theta)_{\theta \in \Theta}$ which forms a cube $\Theta = \left\{\theta = (\theta_1, \dots \theta_d) \in \{\pm 1\}^d\right\}$. We denote $\Lambda(f_{\theta'}, f_\theta)$ the likelihood ratio for n observations. Assume that for any couple of hypotheses $(f_\theta, f_{\theta'})$ such that $\|\theta - \theta'\|_0 = 1$, the likelihood is bounded by*

$$\mathbb{P}_{Y_1, \dots Y_n}\left(\Lambda(f_{\theta'}, f_\theta) \geqslant \beta\right) \geqslant 1 - \alpha,$$

for a suitable $\beta > 0$ *and* $\alpha \in (0,1)$, *then*

$$\inf_{\hat{\theta}\in\Theta} \sup_{\theta\in\Theta} R(f_{\hat{\theta}}, f_\theta) \geqslant \frac{d}{2}(1-\alpha)(\tau \wedge 1).$$

This lemma traduces the amount of difficulty to find the good f_θ when two hypotheses are statistically closed each others (greater than $\beta > 0$) with a probability far enough from 0.

If one expects a simple use of such Lemma when observations are coming from a direct convolution of f with g, there still exists a lot of work to extend the use of such Lemma to the model (3.6). We give in[9] a precise meaning to the likelihood ratio between two hypotheses for the model (3.6) using a conditional argument coupled with the Girsanov formula/ For any measurable ψ of the data Y , we can write

$$\mathbb{E}_{Y\sim f_\theta}[\psi(Y)] = \int_0^1 g(\alpha)\mathbb{E}_{Y\sim f_\theta}[\psi(Y)|\tau=\alpha)]d\alpha = \int_0^1 \mathbb{E}_{Y\sim f_0}[\psi(Y)e^{\langle f_\theta^{-\alpha}, dY\rangle - \|f\|^2/2}|\tau=\alpha)]g(\alpha)d\alpha.$$

There exists an intricate way to simplify this expression : if one considers the null hypothesis f_0 (vanishing f), the law of Y under this null hypothesis does not depend on the random shifts α thus

$$\mathbb{E}_{Y\sim f_\theta}[\psi(Y)] = \mathbb{E}_{Y\sim f_0}\left[\psi(Y)\int_0^1 e^{\langle f_\theta^{-\alpha}, dY\rangle - \|f_\theta\|^2/2)}g(\alpha)d\alpha\right].$$

The likelihood ratio between the two hypotheses $(f_\theta, f_{\theta'})$ is thus defined as

$$\mathbb{E}_{Y\sim f_\theta}[\psi(Y)] = \mathbb{E}_{Y\sim f_{\theta'}}\left[\psi(Y)\underbrace{\frac{\int_0^1 e^{\langle f_\theta^{-\alpha}, dY\rangle - \|f_\theta|^2/2)}g(\alpha)d\alpha}{\int_0^1 e^{\langle f_{\theta'}^{-\alpha}, dY\rangle - \|f_{\theta'}\|^2/2)}g(\alpha)d\alpha}}_{:=\Lambda(f_\theta, f_{\theta'})}\right]. \tag{3.9}$$

We obtain in[9] a lower bound of $\Lambda(f_\theta, f_{\theta'})$ in probability for a particular case of cube which belongs to $B^s_{2,2}$, and the following result holds.

Theorem 3.2.2 *Let* $A > 0$ *and Fourier coefficients of* g *satisfy3.8). If* $\nu > 1/2$ *and* $s > 2\nu + 1$, *there exists* C *that only depends on* A *and* s *such that*

$$\mathcal{R}_n(B^s_{2,2}(A)) \geqslant Cn^{-\frac{2s}{2s+2\nu+1}} \text{ when } n \to +\infty.$$

This result shows that the upper bound obtained using an inverse problem point of view is optimal when $n \to +\infty$ up to a logarithmic factor. Hence, the model (3.6) should be considered as an inverse problem with a known noisy operator (that comes from the random translation whose law g is known). From a technical point of view, the lower bound is obtained using third's order Taylor expansion in the likelihood ratio and concentration results such as Bernstein's inequality.

At last, remark that Fano's Lemma may also be used instead of Assoud's Lemma even if its usage does not seem shorter to obtain a convenient lower bound. Indeed, starting from (3.9), the Kullback-Leibler divergence between f_θ and $f_{\theta'}$ may be written as

$$\mathcal{KL}(f_\theta, f_{\theta'}) = \mathbb{E}_{Y\sim f_\theta} \log\left[\Lambda(f_\theta, f_{\theta'})\right].$$

In the direct convolution situation, the simplest model is written as

$$\forall i \in \{1\dots n\} \quad \forall x \in I \qquad d\tilde{Y}_i(x) = f \star g(x)dx + \epsilon dW_i(x),$$

and the Kullback divergence may be simplified as

$$\widetilde{\mathcal{KL}}(f_\theta, f_{\theta'}) = \mathbb{E}_{Y\sim f_\theta} \log\left[\tilde{\Lambda}(f_\theta, f_{\theta'})\right],$$

where

$$\tilde{\Lambda}(f_\theta, f_{\theta'}) = e^{\langle (f_\theta - f_{\theta'})\star g, dY\rangle + \|f_{\theta'}\star g\|^2/2 - \|f_\theta \star g\|^2/2}.$$

In our random shift framework (3.6), the algebraic simplification « Logarithmic - Exponential » is impossible owing to the likelihood ratio formula given by (3.9). It is yet possible to use Jensen's inequality which yields

$$\mathcal{KL}(f_\theta, f_{\theta'}) \leqslant \log\left[\mathbb{E}_{Y\sim f_\theta} \frac{\int_0^1 e^{\langle f_\theta^{-\alpha}, dY\rangle - \|f_\theta|^2/2)} g(\alpha)d\alpha}{\int_0^1 e^{\langle f_{\theta'}^{-\alpha}, dY\rangle - \|f_{\theta'}\|^2/2)} g(\alpha)d\alpha}\right].$$

But indeed, dealing with this last term is then equivalent to handle a modification of Assouad's Lemma!

Remark 3.2.1 *Even if our upper bound technique is rather classical, this is not the case for the result concerning the lower bound : the key argument is the identification of some elements in $\mathcal{F}$ such that the law of Y is independent from the hidden parameters (here the random shifts). Such a strategy should certainly enable to face some very different problems of lower bound computation following the strategy of invariant hypothesis to the hidden parameters of the model.*

3.2.2 Random deformation through Lie group action

It is possible to describe a natural generalization of the equation (3.6) when deformations are much complex and model geometrical transformations such as rotations, translations, ... in dimension larger thant 1. This framework can have an interest for biomedical imaging censored using Radon transform, or in robotics when a robot take several photos of the same scene but with small variations in the pose of the camera regarding its theoretical position.

In [8], we propose a model which describes such generalization : denote G a Lie group of transformations, *compact and semi-simple*. We are interested in the estimation of $f \in \mathbb{L}^2(G)$ which denotes the Hilbert space of complex valued, square integrable functions on the group G with respect to the Haar measure dg in the following deformable model

$$\forall i \in \{1\dots n\} \quad \forall g \in G \qquad dY_i(g) = f(\tau_i^{-1}.g)dg + \epsilon dW_i(g) \qquad \text{où} \qquad (\tau_i)_{i\in\{1\dots n\}} \text{i.i.d.} \sim h. \tag{3.10}$$

Again, h is the known law of deformations which act on G. Indeed, one can use the same formalism as above following a spectral analysis of the problem. Owing to the compactness of G Peter-Weyl produces for any element of $\mathbb{L}^2(G)$ a Fourier expansion parametrised by the irreducible representations of G which are countable. A Fourier reconstruction formula is then still valid $\mathbb{L}^2$

$$\forall g \in G \qquad f(g) = \sum_{\pi\in\widehat{G}} d_\pi \mathrm{Tr}\left(\pi(g)c_\pi(f)\right),$$

where $\widehat{G}$ denotes the set of irreducible representations of G, d_π is the dimension of the representation π, and $c_\pi(f)$ is the squared matrix which plays the same role as standard Fourier coefficients of f for the eigenvector π of the Laplace Beltrami operator on G with eigenvalue λ_π.

As a consequence, compute an estimation of f is again equivalent to find a suitable way to approach Fourier coefficients on low frequency of f, and threshold the largest ones. To obtain

a suitable frequencies-thresholding , an assumption on the regularity ν of h is necessary when the asymptotic decreasing power s of Fourier coefficients of f is known connue. Let us denote

$$\forall \pi \in \hat{G} \qquad \widehat{c_\pi}(f) = \frac{1}{n}\sum_{j=1}^{n} c_\pi(Y_j)c_\pi(h)^{-1},$$

we build the following estimator $\widehat{f}_n^T$ with the thresholds that omit some large frequencies π :

$$\widehat{f}_n^T = \sum_{\pi \in \hat{G}_T} d_\pi \mathrm{Tr}\left(\pi(g)\widehat{c_\pi}(f)\right).$$

Since the frequency is quantify by the eigenvalue of π, $\hat{G}_T$ is naturally introduced as the set of representations whose eigenvalue λ_π e is lower than T. Moreover, we can also use the generalisation of Sobolev spaces using the above Harmonic analysis if we define for any $A > 0$ the set

$$H_s(A) = \left\{ f \in \mathbb{L}^2(G) \quad | \quad \|f\|_2^2 + \sum_{\pi \in \hat{G}} \lambda_\pi^s d_\pi \|c_\pi(f)\|^2 \leqslant A \right\}.$$

Standard methods of Fourier analysis enable to produce an efficient way to threshold frequencies when s is known and $f \in H_s(A)$. This yields the following theorem.

Theorem 3.2.3 *Assume h known with regularityν and $f \in H_s(A)$ where s is known and such thats $> \dim(G)/2$. Then, for $T_n = n^{\frac{2}{2s+2\nu+\dim G}}$, there exists $K_1 \geqslant 0$ that satisfies*

$$\limsup_{n\to+\infty} \sup_{f\in H_s(A)} n^{\frac{2s}{2s+2\nu+\dim G}} \mathcal{R}(\widehat{f}_n^{T_n}, f) \leqslant K_1.$$

It is still possible to study the likelihood ratios in a similar way as it was already done in (3.9) and then obtain a lower bound for the minimax risk.

Theorem 3.2.4 *Assume h with regularity ν and $s > 2\nu + \dim(G)$, therefore there exists $K_2 \geqslant 0$ such that*

$$\liminf_{n\to+\infty} \inf_{\hat{f}\in\mathbb{L}^2(G)} \sup_{f\in H_s(A)} n^{\frac{2s}{2s+2\nu+\dim G}} \mathcal{R}(\widehat{f}_n^{T_n}, f) \geqslant K_2.$$

Remark 3.2.2 *One should remark that Theorem 3.2.3 is weaker than the upper bound given by Theorem3.2.1. Indeed, we only obtain a non adaptive estimator in Theorem 3.2.3 since s is assumed to be known. This is due to the Fourier thresholding although we reached adaptivity in Theorem 3.2.1 using wavelet expansions. It would be possible to obtain an adaptive estimator for the model (3.10) using the so-called [Lepski, 1991] method, which possesses a very simple principle but with an extensive computational cost in practice.*

3.2.3 Finite horizon approach

The former mathematical studies only provide answers in an asymptotic setting for deformable models. It is however possible to give study this model when the number of curves n remains fixed. In[12], we still study the model :

$$\forall j \in \{1 \dots n\} \quad \forall x \in [0;1] \qquad dY_j(x) = f(x - \tau_j)dx + \epsilon dW_j(x) \quad \text{où} (\tau_j)_{j\in\{1\dots n\}} \text{i.i.d.} \sim g, \quad (3.11)$$

and the quadratic risk keeping an extensive use of Fourier analysis. When k is fixed, we proceed to a preliminary estimation of $c_k(f)$ with simple empirical mean of $(\theta_{j,k})_{j=1\dots n}$. We then use some filtering method through positive filters $(\lambda_k)_{k\in\mathbb{Z}}$ in order to compute $\hat{\theta}(\lambda)$. More precisely,

$$\forall k \in \mathbb{Z} \qquad \hat{\theta}(\lambda)_k = \frac{\lambda_k}{\gamma_k}\frac{1}{n}\sum_{j=1}^{n}\theta_{j,k}.$$

The quadratic risk of estimation may then be decomposed in

$$R(f_{\hat{\theta}(\lambda)}, f) = \underbrace{\sum_{k\in\mathbb{Z}}(\lambda_k - 1)^2|c_k(f)|^2}_{\text{Biais}} + \underbrace{\frac{\epsilon^2}{n}\sum_{k\in\mathbb{Z}}\frac{\lambda_k^2}{|\gamma_k|^2}}_{V_1} + \underbrace{\frac{1}{n}\sum_{k\in\mathbb{Z}}\left[\lambda_k^2|c_k(f)|^2\left(\frac{1}{|\gamma_k|^2}-1\right)\right]}_{V_2}.$$

The bias term is standard but the variance term not since it is composed of two terms : the first one comes from the white noise model in inverse problems and the second traduces the effect of the random translations : we divide (3.7) by γ_k instead of the theoretical unobserved $\tilde{\gamma}_k = \frac{1}{n}\sum_{j=1}^{n} e^{-i2\pi k\tau_j}$ to recover $c_k(f)$ which yields an additional variance term.

In fact, $|c_k(f)|^2$ is unknown an also R thus it is not possible to optimize the choice of λ to obtain a correct inference on f. However, it is possible to build an estimation $|\hat{\Theta}_k|^2$ of $|c_k(f)|^2$ and then follow the *Unbiased Risk Estimation*) in our framework. We first define for any $\alpha \in [0;1]$

$$U_\alpha(Y,\lambda) = \sum_{k\in\mathbb{Z}}(\lambda_k^2 - 2\lambda_k)|\gamma_k|^{-2}|\hat{\Theta}_k|^2 + \frac{\epsilon^2}{n}\sum_{k\in\mathbb{Z}}\lambda_k^2|\gamma_k|^{-2} + \alpha\frac{\log^2 n}{n}\sum_{k\in\mathbb{Z}}\lambda_k^2|\gamma_k|^{-4}|\hat{\Theta}_k|^2.$$

If we consider the restrictive class of symmetric and monotones in $|k|$ filters :

$$\Lambda_{mon} := \left\{\lambda = (\lambda_k)_{k\in\mathbb{Z}} : \lambda_k = \lambda_{-k}, \sum_{k\in\mathbb{Z}}\lambda_k^2 < +\infty,\ 1 \geqslant \lambda_0 \geqslant \dots \geqslant \lambda_m \geqslant \dots \geqslant 0\right\},$$

it is possible to compute the « optimal » filters

$$\hat{\lambda}_\alpha = \arg\min_{\lambda\in\Lambda_{mon}} U_\alpha(Y,\lambda).$$

This optimal filter $\hat{\theta}(\hat{\lambda}_\alpha)$ then satisfies an oracle inequality according to the next Theorem.

Theorem 3.2.5 *Assume that Fourier coefficients of g satisfy the property (3.8), then there exists $\gamma_1 \in (0,1)$ such that for all $\gamma \in (0,\gamma_1)$,*

$$\mathbb{E}_\theta\|\hat{\theta}(\hat{\lambda}_\alpha) - c_\cdot(f)\|^2 \quad\leqslant\quad (1+h_{\gamma,n})\inf_{\lambda\in\Lambda_{mon}}\left[R(f_{\hat{\theta}(\lambda)}, f) + \alpha\frac{\log^2 n}{n}\sum_{k\in\mathbb{Z}}\lambda_k^2|\gamma_k|^{-2}|c_k(f)|^2\right] + \Gamma_{\gamma,n,\epsilon^2}(c(f),\alpha)$$

where $h_{\gamma,n} \to 0$ when $\gamma \to 0$ and $n \to +\infty$, and $\Gamma_{\gamma,n,\epsilon^2}(c(f),\alpha)$ is an explicit function of (γ, n, ϵ^2) and $(c(f),\alpha)$.

The description of the map Γ is rather technical and we refer to [12] for more details. This function Γ possesses essentially two kinds of terms : one term has a decreasing property of order ϵ^2/n and the other one of order $\log^2 n/n$. One sould also note that α traduces a balance in the Signal to Noise Ratio (see the numerical study in [12]) and should be chosen near 0 for large ϵ and in the opposite case quite large when the SNR increases.

3.3 Deformable model with unknown deformation law

3.3.1 Statements

Paragraphs of Section 3.2 described some result on the estimation of f when the law of deformations is known for problems such as (3.6) or (3.10) and concerning the asymptotic $n \to +\infty$ and when the noise level ϵ is a fixed parameter. In some cases, the knowledge of g may be satisfactory since in some image processing problems, some calibration of censor may be done to estimate g before real measurements on the dataset. This may be the case for instance if one considers tomographic images obtained through Randon transform if one decides to estimate g on preliminary patients. This may also be the case for the calibration of a camera which is moving around a theoretical position. However, this framework may not be suitable for other practical examples when we do not have any control on a preprocessing step. Hence, some works should also be developed in the case where g is unknown to estimate f.

We can number at least two motivations for this study. We may be interested in the estimation of the way data are generated and thus we would make some descriptive statistics in such models. We may also try to find the deformation parameters which are unobserved and then obtain an estimation of the signal f himself. Loosely speaking, if one observes

$$\forall j \in \{1 \dots n\} \quad \forall x \in [0;1] \qquad dY_j(x) = f(x - \tau_j)dx + \epsilon dW_j(x) \quad \text{où} (\tau_j)_{j\in\{1\dots n\}} \text{i.i.d.} \sim g,$$

the simplest method to compute f should consider an estimation procedure of the deformation parameters $(\hat{\tau}_j)_{j\in\{1\dots n\}}$, and then invert these deformations $\hat{\tau}_j$ on each signal Y_j in order to estimate f by a simple empirical mean :

$$\hat{f}_n(.) = \frac{1}{n}\sum_{j=1}^{n} Y_j(. + \hat{\tau}_j). \tag{3.12}$$

In the sequel, we are interested by the two following questions :
- Is it possible to recover the deformation parameters ?
- Is it possible to recover f without any knowledge on g ?

Remark 3.3.1 *One should precise that our asymptotic study is not concerned by the semi-parametric problem when curves $(Y_i)_{i\in\{1\dots n\}}$ are observed on a grid which may more and more accurate. One should refer to several recent works of [Gamboa et al., 2007b], [Bigot et al., 2009] and [Vimond, 2010]) in this isssue. Remark that formally, make an asymptotic study when the sampling frequency of the grid is growing is equivalent to an asymptotic study where $\epsilon \to 0$ in model (3.6).*

3.3.2 Frechet mean to estimate f

One may consider a global non parametric estimation of f using Fréchet mean of random variables $Z_1, \dots, Z_n$ which do not belong to a vectorial space V. This fact is consistent with the remark that Z, Z' may be considered as identical in our model if one can find a transformation in a group H which send exactly Z on Z'. In [Frechet, 1948], the euclidean mean is extended to general metric spaces through an implicit criterion : consider a distance d defined on a manifold $\mathcal{M}$, the Fréchet mean of n observations $(Z_i)_{i\in\{1\dots n\}}$ of $\mathcal{M}$ is given by

$$\hat{Z}_n^F = \arg\min_{Z\in\mathcal{M}} \frac{1}{n}\sum_{m=1}^{n} d^2(Z, Z_m).$$

In our framework of randomly shifted curves, $H = \mathbb{R}$ is the group of translation acting on $f \in L^2([0,1])$ by

$$\tau \cdot f(x) = f(x+\tau), \qquad \text{for } x \in [0,1] \text{ and } \tau \in H.$$

Let be given n observations $Y_1, \dots, Y_n$ through (3.6), the Fréchet mean under the action of H is then

$$\hat{f}_n^F = \arg \min_{f \in L^2([0,1])} \frac{1}{n} \sum_{m=1}^{n} \min_{\tau_m \in \mathbb{R}^+} \int_0^1 |f(x) - Y_m(x+\tau_m)|^2 dx.$$

If one considers the Fourier coefficients of the data (denoted $\theta_{m,\ell}$ at frequence ℓ for observation m), and if we use ℓ_0 as a threshold frequency, the estimation $(\hat{\theta}_k)_{-\ell_0 \leqslant k \leqslant \ell_0}$ is then

$$(\hat{\theta}_{-\ell_0}, \dots, \hat{\theta}_{\ell_0}) = \arg \min_{(\theta_{-\ell_0}, \dots, \theta_{\ell_0}) \in \mathbb{R}^{2\ell_0+1}} \frac{1}{n} \sum_{m=1}^{n} \min_{\tau_m \in \mathbb{R}} \sum_{|\ell| \leqslant \ell_0} |\theta_{m,\ell} e^{2i\ell\pi\tau_m} - \theta_\ell|^2. \tag{3.13}$$

Thus, Fréchet mean is then obtained by Fourier reconstruction $\hat{f}^F_{n,\ell_0}(x) = \sum_{|\ell| \leqslant \ell_0} \hat{\theta}_\ell e^{-2i\ell\pi x}$. At last, remark that (3.13) possesses an explicit solution $\hat{\theta}_\ell = \frac{1}{n} \sum_{m=1}^{n} \theta_{m,\ell} e^{2i\ell\pi\hat{\tau}_m}$, and thus

$$(\hat{\tau}_1, \dots, \hat{\tau}_n) = \arg \min_{(\tau_1, \dots, \tau_n) \in \mathbb{R}^n} \underbrace{\frac{1}{n} \sum_{m=1}^{n} \sum_{|\ell| \leqslant \ell_0} \left| \theta_{m,\ell} e^{2i\ell\pi\tau_m} - \frac{1}{n} \sum_{q=1}^{n} \theta_{q,\ell} e^{2i\ell\pi\tau_q} \right|^2}_{:= M_n(\tau_1, \dots \tau_n)}. \tag{3.14}$$

To sum up, the Fréchet mean computation is equivalent to the minimisation of the criterion defined in equation (3.14), which may be solved by a gradient descent algorithm.

3.3.3 Estimation of the parameter of deformations

Remind that (3.6) is equivalent in the Fourier basis to

$$\theta_{m,\ell} = c_\ell(f) e^{-i2\pi\ell\tau_m^*} + \epsilon z_{\ell,m}, \ \ell \in \mathbb{Z} \text{ for } m = 1, \dots, n, \tag{3.15}$$

where $z_{\ell,m}$ are i.i.d. $\mathcal{N}_{\mathbb{C}}(0,1)$ and $\tau_m^*, m = 1, \dots, n$ are the true translation parameters sampled with the unknown law g. Problem (3.15) is clearly not uniquely identifiable since for any $\tau_0 \in \mathbb{R}$, one may consider $\theta_\ell e^{i2\pi\ell\tau_0}$ instead of θ_ℓ and $\tau_m^* - \tau_0$ instead of τ_m^* without any modification of the data. We thus introduce the two following identifiability conditions :

Assumption 6 (H_g) *g is centered and compactly supported by $\mathcal{T} = [-\frac{1}{4}, \frac{1}{4}]$.*

Assumption 7 (H_f) *f is such that $c_1(f) \neq 0$.*

From assumption (H_g), we restrict our estimations to a set of n empirically centered parameters $(\hat{\tau}_1, \dots, \hat{\tau}_n)$. We thus introduce

$$\overline{\mathcal{T}}_n = \{(\tau_1, \dots, \tau_n) \in \mathcal{T}^n \text{ tels que } \sum_{m=1}^{n} \tau_m = 0\},$$

where the frequency threshold ℓ_0 remains fixed. We then look at $\tau = (\tau_1, \dots, \tau_n) \in \overline{\mathcal{T}}_n$ which optimises $M_n(\tau)$ since this is the only requirement to build the Fréchet mean $\hat{f}_n^F$. One can then expect to recover the deformation parameters.

Theorem 3.3.1 *Assume that* (H_g) *and* (H_f) *are in force, and define*

$$\hat{\tau} = \arg\min_{\tau \in \overline{\mathcal{T}}_n} M_n(\tau).$$

For any $t > 0$*, one has*

$$\mathbb{P}\left(\frac{1}{n}\sum_{m=2}^{n}(\hat{\tau}_m - \tau_m^*)^2 \geqslant C(f, \ell_0, \epsilon, n, t, g)\right) \leqslant 3\exp(-t), \tag{3.16}$$

where $C(f, \ell_0, \epsilon, n, t, g) = 4\max\left[C_1(f, \ell_0)\left(\sqrt{C_2(\epsilon, n, \ell_0, t)} + C_2(\epsilon, n, \ell_0, t)\right), C_3(t, n, g)\right]$*. Note that* $C_1(f, \ell_0)$ *is a non negative constant that only depends on* f *and the threshold* ℓ_0*, while*

$$C_2(\epsilon, n, \ell_0, t) = \epsilon^2(2\ell_0 + 1) + 2\epsilon^2\sqrt{\frac{2\ell_0 + 1}{n}t} + 2\frac{\epsilon^2}{n}t,$$

and

$$C_3(t, n, g) = \left(\sqrt{2\epsilon_g^2\frac{t}{n}} + \frac{t}{12n}\right)^2 \text{ où } \epsilon_g^2 = \int_{\mathcal{T}} \tau^2 g(\tau)d\tau.$$

Theorem 3.3.1 provides then an upper bound in probability to the accuracy of the estimation of deformation parameters $\hat{\tau}$ comparing to the true ones $\tau_m^*, m = 2, \ldots, n$. The minimum of $M_n(\tau)$ is computed on $\overline{\mathcal{T}}_n$, thus $\hat{\tau}_1 = -\sum_{m=2}^{n}\hat{\tau}_m$. Hence, when $n \to +\infty$, $C(f, \ell_0, \epsilon, n, t, g)$ which is used in (3.16) converges towards $4C_1(f, \ell_0)\left(\epsilon^2(2\ell_0 + 1) + \epsilon\sqrt{2\ell_0 + 1}\right)$ and we cannot obtain with this bound a consistent estimation. Indeed $\lim_{n\to+\infty}\frac{1}{n}\sum_{m=2}^{n}(\hat{\tau}_m - \tau_m^*)^2 = 0$ in probability seems impossible and (3.16) rather suggests that there exists $C > 0$ such that $\frac{1}{n}\sum_{m=2}^{n}(\hat{\tau}_m - \tau_m^*)^2 > C\epsilon^2(2\ell_0 + 1)$ with a positivie probability. Thus, the accuracy of $\hat{\tau}$ should depend on the noise level ϵ^2 and the threshold ℓ_0.

3.3.4 Lower bound of reconstruction

En supposant que f est de classe $\mathcal{C}^1([0, 1])$, il est possible de donner une borne inférieure de reconstruction des paramètres de déformation qui dit en substance que (3.16) est presque optimale et que si le niveau du bruit ϵ est maintenu constant, alors il n'est pas possible d'estimer les $(\tau_m^*)_{m=1\ldots n}$ même en répétant les observations $(n \mapsto +\infty)$. Plus précisément, on suppose

Assumption 8 ($\tilde{H}_g$) *The unknown law* g *is compactly supported on* $\mathcal{T}$ *with* $\lim_{\tau\to\inf\mathcal{T}} g(\tau) = \lim_{\tau\to\sup\mathcal{T}} g(\tau) = 0$.

Assumption 9 ($\tilde{H}_f$) f *satisfies* $\|f'\|_2^2 = \sum_{\ell\in\mathbb{Z}}(2\pi\ell)^2|c_\ell(f)|^2 < +\infty$.

It is then possible to show the following result.

Theorem 3.3.2 *Let* $X = (\theta_{m,\ell})_{\ell\in\mathbb{Z}, m=1,\ldots,n}$ *be the set of Fourier coefficients observed in* $\mathbb{X} = \ell^2(\mathbb{Z})^{\otimes n}$ *and denote* $\hat{\tau}^n = \hat{\tau}^n(X) \in \mathbb{X}$ *any measurable function of* X*. Assume that* $(\tilde{H}_f)$ *and* $(\tilde{H}_g)$ *hold, then we have*

$$\mathbb{E}\left(\frac{1}{n}\sum_{m=1}^{n}(\hat{\tau}_m^n - \tau_m^*)^2\right) \geqslant \frac{\epsilon^2}{\|f'\|_2^2 + \epsilon^2 I(g)},$$

where $I(g)$ *is the Fisher information*

$$I(g) = \int_{\mathcal{T}}\left(\frac{\partial}{\partial\tau}\log g(\tau)\right)^2 g(\tau)d\tau.$$

The proof of Theorem 3.3.2 uses a van Trees inequality which can be viewed as a Bayesian Cramer-Rao inequality bayésienne. When $n \to +\infty$, $\mathbb{E}\left(\frac{1}{n}\sum_{m=1}^{n}(\hat{\tau}_m^n - \tau_m^*)^2\right)$ cannot converge to 0 and it explains the upper bound obtained by Theorem 3.3.1. Note also that it is possible to weaken the assumption $f \in \mathcal{C}^1([0,1])$ by considering estimators $\hat{\tau}^{n,\ell_0}$ built from $\theta_{m,\ell}$ for $m = 1,\dots,n$ et $|\ell| \leqslant \ell_0$ in model (3.15). In this case, the lower bound is then

$$\mathbb{E}\left(\frac{1}{n}\sum_{m=1}^{n}(\hat{\tau}_m^{n,\ell_0} - \tau_m^*)^2\right) \geqslant \frac{\epsilon^2}{\sum_{|\ell|\leqslant\ell_0}(2\pi\ell)^2|\theta_\ell|^2 + \epsilon^2\int_{\mathcal{T}}\left(\frac{\partial}{\partial\tau}\log g(\tau)\right)^2 g(\tau)d\tau}.$$

3.3.5 Mean pattern recognition with deformable models

The Fréchet mean approach described above may be extended to the case of images which are corrupted by general deformations that belong to a Lie group. We describe in [11] a model of elastic deformations which are coming from flows of diffeomorphism described by (3.3). If we consider the set of grey levelled images defined on $\Omega \subset \mathbb{R}^2$, a pattern I is just an application $I : \Omega \longmapsto \mathbb{R}$. We aim to interpret our Fréchet mean as a simple M-estimator (see for instance[Van der Waart, 1998] for further details on these estimators) based on a special contrast.

First, we propose to use a parametrisation of diffeomorphisms using the approach developed by [Trouvé and Younes, 2005] for homogeneous vector fields in equation (3.3). Without loss of generality, we fix $\Omega = [0,1]^2$ and impose a parametric structure on $v : [0,1]^2 \mapsto \mathbb{R}^2$ that satisfies $v_{\partial[0,1]^2} = 0$. Hence, if $(e_1,\dots e_K)$ is a finite family of basis functions from $[0;1]^2$ to $\mathbb{R}^2$, vanishing on $\partial[0,1]^2$, we obtain a random vector field v_a by a simple generation of 2K coefficients $(a_1^1,\dots a_K^1) \times (a_1^2,\dots a_K^2)$ for which

$$\begin{cases} v_a^1 = \sum_{j=1}^{K} a_j^1 e_k^1 \\ v_a^2 = \sum_{j=1}^{K} a_j^2 e_k^2 \end{cases}$$

Let be given v_a, a random diffeomorphism is then obtained using simply solution at time 1 of (3.3) which will be denoted $\Phi_{v_a}^1$. Such construction can be extended to a random image deformable model. If P_A is a compactly supported law in $[-A,A]$ with $A > 0$ and if K denotes any positive integer, we define our model of randomly warped image as

$$\forall p \in [0,1]^2 \qquad I_{\epsilon,a} = I^* \circ \Phi_{v_a}^1(p) + \epsilon(p), \tag{3.17}$$

where ϵ is an additive noise independent from the coefficients $a \sim P_A^{\otimes 2K}$. We consider now n realisations sampled independently of (3.17) which are denoted I_{ϵ_i,a_i}. V_A will be the set of reachable vector fields with coefficients that live in $[-A,A]^{2K}$. For a given image Z defined on $[0,1]^2$, the following contrast function $f(a,\epsilon,Z)$ uses a pixel discretisation $\mathcal{P}$ of $[0,1]^2$:

$$f(a,\epsilon,Z) = \min_{v\in V_A} |I_{\epsilon,a} - Z \circ \Phi_v^1|_{\mathcal{P}}^2.$$

Hence, f corresponds to the minimal $\ell^2(\mathcal{P})$ cost when Z is matched on $I_{\epsilon,a}$ using V_A. The mean contrast function is then

$$F(Z) = \int f(a,\epsilon,Z)dbP(a,\epsilon),$$

and the Fréchet intrinsec mean $I_{\epsilon,a}$ is $Q^* = \arg\min_{Z\epsilon\mathcal{Z}} F(Z)$. If $\mathbb{P}_n$ denotes the empirical measure of the data, it is also possible to define the empirical contrast by

$$F_n(Z) = \int f(a,\epsilon,Z)d\mathbb{P}_n(a,\epsilon) = \frac{1}{n}\sum_{j=1}^{n}\min_{v_j\in V_A}|I_{\epsilon_j,a_j} - Z\circ\Phi_{v_j}^1|_{\mathcal{P}}^2. \tag{3.18}$$

The Fréchet mean of the data as the minimizer $\hat{Q}_n = \arg\min_{Z\in\mathcal{Z}} F_n(Z)$. The main advantage is that one can compute $\hat{Q}_n$ although Q^* is intractable since the law of the deformation is unknown.

Of course, some minimization procedure such as equation (3.18) may yield very different estimators from I^* and the contrast F_n should be regularized. In [11], we then add to F_n a penalisation term that aims to control the smoothness of Z as well as the amount of deformation which is allowed to warp images. It is then possible to obtain a.s. convergence properties of these Fréchet mean $\hat{Q}_n$ towards Q^*.

One should remark that we do not know at once if I^* belongs to Q^*, thus our method provides only a very partial answer on this model.

3.4 Numerical results

3.4.1 Randomly shifted curve model

We first provide few numerical experiments on the problem recovering f when data are issued from the random shift model. Four test functions f are studied (see Figures 3.5(a)-3.8(a)) and we observe $n = 200$ noisy and randomly shifted curves using a Laplace law whose density is given by $g(x) = \frac{1}{\sqrt{2}\sigma}\exp\left(-\sqrt{2}\frac{|x|}{\sigma}\right)$ with $\sigma = 0.1$. A sub-sample of 10 curves are given in Figures 3.5(b)- 3.8(b) for each mean pattern. At last, we provide a result of simple averaging for the estimation of f in Figures 3.5(c)- 3.8(c). We immediately remark the poor performance which build estimate the convolution by g and is far from being satisfactory.

Fourier coefficients of g are given by $\gamma_\ell = \frac{1}{1+2\sigma^2\pi^2\ell^2}$, which corresponds to an order of inverse problem $\nu = 2$. In Figures 3.5(d)(e) -3.8(d)(e), we provide the estimation of f using our inverse problem point of view : two thresholding methods are tested $\hat{f}_n^H$ and are described in paragraph 3.2.1 (these two methods are much more detailed in [9]). At last, the estimations obtained without the knowledge of g are given in Figures 3.5(f) -3.8(f) using the Fréchet mean described in paragraph 3.3.2 (estimation $\hat{f}_n^F$). We immediately remark the efficiency of the two last methods, especially the one even when g is unknown and when one should approach its Fourier coefficient.

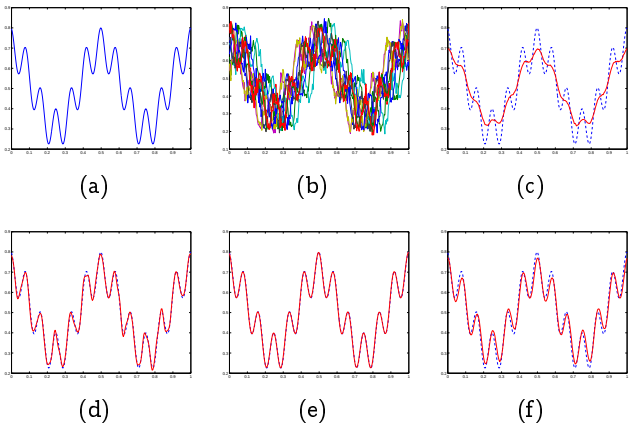


FIGURE 3.5 – Function "Wave". (a) True function f, (b) Sample of 10 curves among $n = 200$, (c) Empirical mean, Deconvolution (d) $\hat{f}_{n,1}^H$ and (e) $\hat{f}_{n,2}^H$, (f) Fréchet mean

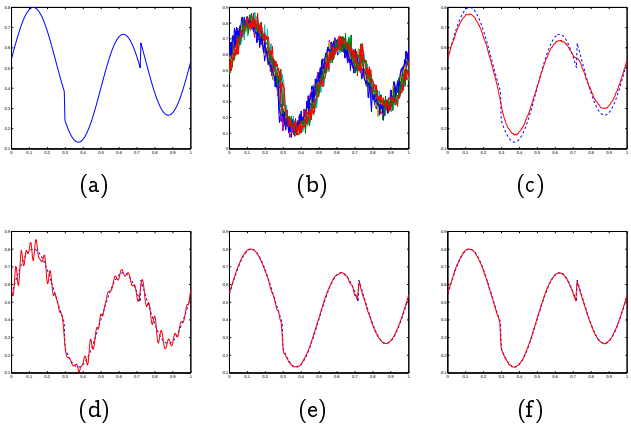


FIGURE 3.6 – Function HeaviSine. (a) True function f, (b) Sample of 10 curves among $n = 200$, (c) Empirical mean, Deconvolution (d) $\widehat{f}^{H}_{n,1}$ and (e) $\widehat{f}^{H}_{n,2}$, (f) Fréchet mean

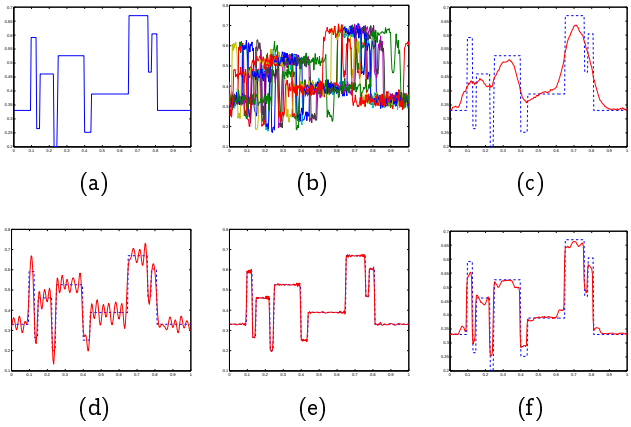


FIGURE 3.7 – Function Blocks. (a) True function f, (b) Sample of 10 curves among $n = 200$, (c) Empirical mean, Deconvolution (d) $\widehat{f}^{H}_{n,1}$ and (e) $\widehat{f}^{H}_{n,2}$, (f) Fréchet mean

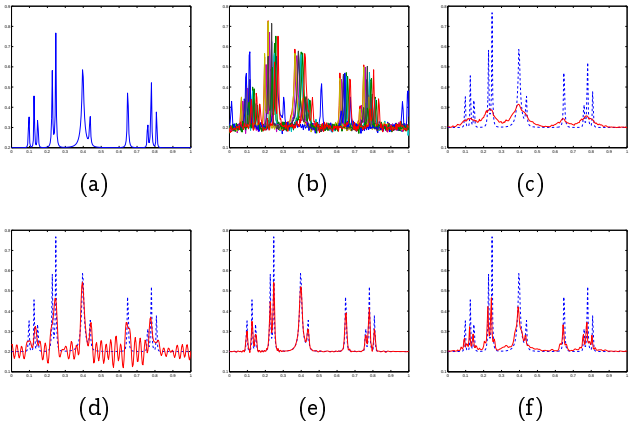


FIGURE 3.8 – Function Bumps. (a) True function f, (b) Sample of 10 curves among $n = 200$, (c) Empirical mean, Deconvolution (d) $\widehat{f}^{H}_{n,1}$ and (e) $\widehat{f}^{H}_{n,2}$, (f) Fréchet mean

3.4.2 Fréchet mean of images

We provide now some numerical experiments on the approach described in paragraph 3.3.5. For the classical Lena image, Figure 3.9 illustrates some deformations enabled by the model of

differential flow of diffeomorphisms where a_k are uniformly sampled on $[-A, A]$. The amount of deformation is defined through the size of A and the use B-spline enables to localise the deformation effects.

FIGURE 3.9 – Random deformation of Lena with $A = 0.1$ et $A = 0.5$.

At last, we show the warping result using Fréchet means coupled with the action of diffeomorphisms flows and compare to the euclidean mean on two famous dataset : the Mnist one of handwritten digits (see Figure 3.10 for the effect of the algorithm [11] on digit "2") and the Olivetti faces one (see Figure 3.11 for some examples of faces warping).

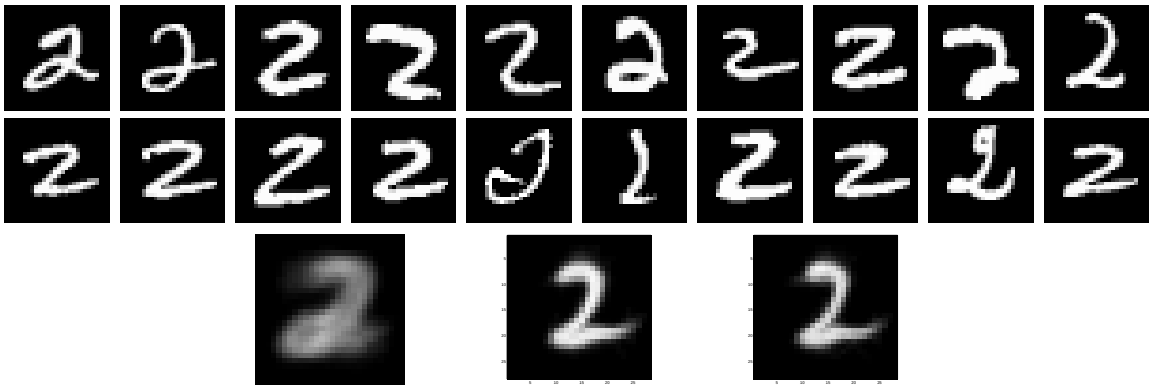


FIGURE 3.10 – Empirical mean (bottom left), first iteration mean $Z_\star^{(1)}$ and third one $Z^{(3)}$ of the algorithm described in [11].

Our method also permits to develop a clustering algorithm for warped images. We refer to [11] for further details on these experiments.

3.5 Further developments

3.5.1 Shape constrained regression

The estimator of constrained monotone regression appears to be really efficient regarding other estimators found on this topic. It would be useful to develop a software for a larger diffusion.

Moreover, one may wonder if a similar approach using diffeomorphisms derived from vector fields may be extended to the case of convex or concave regression which is another shape constrained regression problem which is commonly encountered in some practical problems. This may have some interests in finance [Ait-Sahalia and Duarte, 2003] for stock-options pricing, in econoics [Allon et al., 2007] where the supply and demand are supposed concave functions, or in biology [Ratkowsky, 1983], in respone surface estimation for optimisation tasks

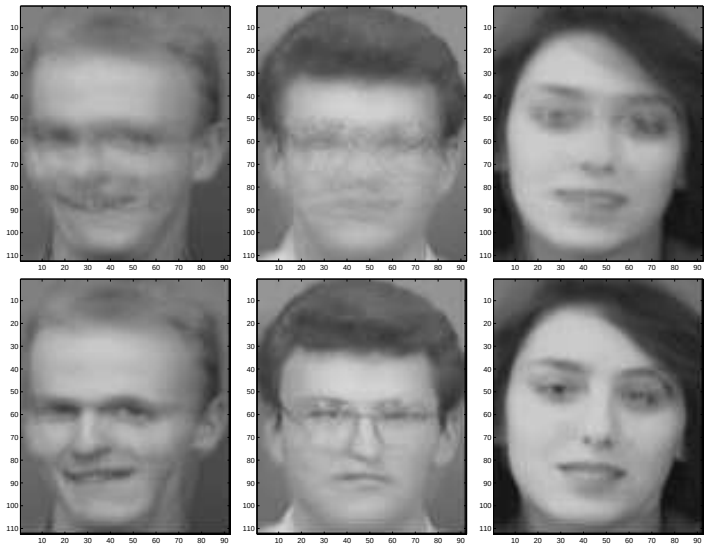


FIGURE 3.11 – Examples of Fréchet mean obtained on 3 faces of the Olivetti database. First line : empirical mean, second line : Fréchet mean.

[Hoffmann et al., 2006] ... If the estimation should be considered on $\mathbb{R}^d$, it would be convenient to consider a time evolution initialized with $\phi_0(x) = |x|^2$ and $\frac{d\phi_t}{dt} = v_t(\phi_t)$ where v_t would preserve the convexity all along the time evolution.

At last, from a theoretical point of view, the convergence obtained in Theorem 3.1.2 may certainly refined. Indeed, we prove in [10] only a convergence in probability of an estimator $(v_t^{n,\lambda})$ to the optimal theoretical vector field v_t by M-estimation techniques. We should consider now some more precise results of M-estimation for infinite dimensional objects following Donsker classes results (see for instance[van der Vaart and Wellner, 1996]). Their main interest would be to obtain statistical testing procedures for shape hypothesis.

3.5.2 Bayesian estimation with unknown operator

In order to simplify the problem, I will limit this paragraph to the randomly shifted curves model given by (3.6) :

$$\forall j \in \{1 \dots n\} \quad \forall x \in [0;1] \qquad dY_j(x) = f(x - \tau_j)dx + \epsilon dW_j(x) \quad \text{où} (\tau_j)_{j\in\{1\dots n\}} \text{i.i.d.} \sim g,$$

where g is *unknown*. We have seen in Theorem 3.3.2 that we cannot recover the deformation parameters τ_j without any asymptotic assumption on the noise level $\epsilon \to 0$, which significantly harms the chance of consistency procedures as Fréchet mean estimation of f described by (3.13) and (3.14). However, it does not seem impossible to approach f without any individual deformation parameter estimation, using in this view a Bayesian point of view.

This is for instance the approach used in [Allassonière et al., 2007] and [Allassonière et al., 2009] where the unknown law g is assumed described by several parameters which are estimated by Bayesian statistics. Note that their work does not assert the statistical consistency when $n \to +\infty$ although it would be a very interesting and instructive problem for a generalization to non parametric family of law g.

Indeed, if one refers to pioneered works of[Ibragimov and Has′minskiĭ, 1981, Le Cam, 1973] on Bayesian consistency, there is some sufficient conditions to ensure such asymptotic good behaviour. Roughly speaking, and in a parametric setting at start, assume $(X_1, \dots X_n)$ to be

i.i.d. realisations of P_{θ_0} where $\theta_0 \in \Theta \subset \mathbb{R}^d$ is unknown, Bayesian estimator of θ_0 depends on a prior q which is a probabilistic distribution of Θ. Assume moreover that each $(P_\theta)_{\theta\in\Theta}$ possesses some density $(p_\theta)_{\theta\in\Theta}$ with respect to a common reference measure on $\mathcal{X}$ (space where each X_i is living), Bayesian estimator is then defined as

$$\hat{\theta}_n^B = \arg\min_{\theta\in\Theta} \int_\Theta L(u-\theta)p_\theta(X_1,\dots,X_n)q(u)du,$$

where L is a loss function, vanishing at $0_{\mathbb{R}^d}$ (indeed $L(x) = |x|^p$ for any $p > 0$ is the typical case). The striking Theorem 5.2 of chapter 1 in [Ibragimov and Has′minskiĭ, 1981] asserts consistency provided the likelihood ratios between two hypothesis is sufficiently regular when θ varies, becomes small when θ is far from θ_0 and the prior q is a continuous strictly positive density on Θ. Moreover, one should note that this is the smoothness of this likelihood ratio arount θ_0 which describes the convergence rate of $\hat{\theta}_n^B$: the more flat is the likelihood ratio, the less rate are fasts. This smoothness on likelihood ratio may be described by separability conditions on Hellinger distances or Kullback-Leibler divergences [1]. At last, this Bayesian estimator may be interpreted as a special case of perturbed Laplace method for integrals where the main mode of integrands is convergent and the posterior distribution of θ given $(X_1,\dots X_n)$ becomes a Gaussian distribution whose variance depends on the Fisher information at θ_0 divided by the convergence rate ($\sqrt{n}$ usually) : these results are described by Bernstein-von Mises Theorem (described in [Le Cam, 1973] or [Van der Waart, 1998]).

In a non parametric framework, the situation seems more intricate. Some recent advances extend the parametric results of [Le Cam, 1973] and [Ibragimov and Has′minskiĭ, 1981] on posterior distributions using some covering arguments and uniform lower bound of Hellinger distances related to other distance intrinsic distance on Θ. We may cite, among a large amount of litterature, the works of [Ghosal et al., 2008, Ghosal, 2000, Rousseau, 2010]. The main idea is to construct growing sieves with n and entropy controls.

Of course in our model (3.6), the additional problem already present in the frequentist approach, is still here since observations depend on the hidden parameters $(\tau_i)_{i=1\dots n}$ sampled following g. If one consider the simplest case of estimating only one Fourier coefficient θ_0 of f, the observations are given by

$$\forall i \in \{1\dots n\} \qquad \theta_i = e^{2i\pi\tau_i}\theta_0 + \epsilon_i.$$

If q denotes a prior on Θ and r a prior on the $\mathcal{L}^1(\mathcal{S}^1)$ (space of density on the one dimensional sphere parametrised by $e^{2i\pi\tau}$), it is possible to model a Bayesian estimation problem for (θ_0, g) :

$$(\hat{\theta}_n^B, \hat{g}_n^B) = \arg\min_{\theta\in\Theta, g\in\mathcal{L}^1(\mathcal{S}^1)} \int_{\Theta\times\mathcal{L}^1(\mathcal{S}^1)} L(u-\theta; v-g)p_{\theta,g}(X_1,\dots X_n)I_{u,v}(X_1,\dots,X_n)dq(u)dr(v)$$

where $I_{u,v}(X_1,\dots,X_n)$ is the likelihood ratio with given priors q, r. Some method to face this infinite mixture problem may adapt the approach of [Rousseau, 2010] (for instance).

A first step to obtain good behaviour of the Bayesian posterior and estimators should be to establish sufficients condition that ensure identifiability of the model, both for the unknown f and the unknown density g. Such result could be obtained using the link between the total variation distance between $d_{VT}(\mathbb{P}_{f,g}, \mathbb{P}_{\tilde{f},\tilde{g}})$ and the Laplace transform $\mathcal{L}(g-\tilde{g})$ and the Fourier expansion of f and $\tilde{f}$. Next, it is necessary to bound from above the covering numbers of the law $\mathbb{P}_{f,g}$ with respect to the Hellinger distance (or Kullback Leibler or total variation distance) but this may not be derived from a "standard" inequality between these distances and some distance on $f, \tilde{f}, g, \tilde{g}$. It is thus necessary to work on the infinite mixture of Gaussian laws.

1. Indeed, only Kullback-Leibler divergence is necessary, which is a slightly weaker condition.

3.5.3 Randomly shifted Poissonian noise

In [16], we extend our asymptotic study on randomly shifted curves corrupted with a white noise to the case where observations are issued from a counting process which is modelled by a Poisson process whose inhomogeneous intensity λ is unknown and should be estimated. Observations are then obtained with several counting processes defined on $[0,1]$ $N^1, \dots N^n$ with intensities $\lambda(.-\tau_1), \dots, \lambda(.-\tau_n)$. Of course, $(\tau_i)_{i=1\dots n}$ are i.i.d. observations sampled with g which is assumed to be known at the moment. Such model may describe a random phenomenon observed on Chip-Seq datasets. We prove in [16] that one may recover λ with some convergence rate similar to the one obtained in the white noise model. More precisely, we obtain the following theorem.

Theorem 3.5.1 *Assume g known and that satisfies the inverse problem order ν hypothesis (3.8), we define moreover*

$$\Lambda_0 = \left\{ \lambda \in L^2([0,1]);\ \lambda(t) \geqslant 0 \text{ pour tout } t \in [0,1] \right\}.$$

Let $A > 0$ and assume that λ has a smoothness parameter s such that $s > 2\nu + 1$. Hence, there exists $C_0 > 0$ (independent of n) such that for n sufficiently large

$$\inf_{\hat{\lambda}_n} \sup_{\lambda \in B^s_{2,2}(A) \bigcap \Lambda_0} \mathcal{R}(\hat{\lambda}_n, \lambda) \geqslant C_0 n^{-\frac{2s}{2s+2\nu+1}},$$

where the minimum is computed over all estimators $\hat{\lambda}_n \in \Lambda_0$ (i.e. measurable functions of processes N^i, $i = 1, \dots, n$ with values in Λ_0).

The proof of this lower bound is again dependent on a precise control of the likelihood ratio between hypothesis λ and $\lambda + h$ coupled with an Assouad's like lemma. This likelihood ratio may be written using a Girsanov formula for Poisson processes using again a « null »hypothesis, which is an hypothesis invariant by any random shift. Here, the null hypothesis corresponds to a constant intensity ρ defined on $[0,1]$. If (λ_1, λ_2) are two intensities such that $\lambda_1 \geqslant \rho > 0$ and $\lambda_2 \geqslant \rho > 0$, and for an observed counting process N, the likelihood is

$$\Lambda(\lambda_1, \lambda_2)(N) = \frac{\int_0^1 \exp\left[-\int_0^1 \mu_1(t-\alpha)dt + \int_0^1 \log\left(1 + \frac{\mu_1(t-\alpha)}{\rho}\right) dN_t \right] g(\alpha)d\alpha)}{\int_0^1 \exp\left[-\int_0^1 \mu_2(t-\alpha)dt + \int_0^1 \log\left(1 + \frac{\mu_2(t-\alpha)}{\rho}\right) dN_t \right] g(\alpha)d\alpha)}$$

where $\mu_1 = \lambda_1 - \rho$ and $\mu_2 = \lambda_2 - \rho$.

Finally, we propose an adaptive estimator $\hat{\lambda}^h_n$ based on hard thresholding methods which is asymptotically optimal up to a logarithmic term $\mathcal{O}\left(\left(\frac{\log n}{n}\right)^{\frac{2s}{2s+2\nu+1}}\right)$. This estimator is precisely described in[16].

At last, we should note that such model seems to be the good statistical framework for Chip-Seq data which are counting processes censored with some rigid geographical deformation effects on the DNA. Hence, experimenters actually use some warping procedures before analysing the data. The kind of estimator we propose could then deal automatically with such problem.

3.5.4 Statistical testing problems

At last, it would be of interest to extensively use the likelihood ratio structure presented above to build statistical tests on the hypothesis : two sampled curves follow the same randomly

shifted white noise model or not. This may be possible following the last works on these models, and maybe could be extended to a Poissonian noise instead of Gaussian one. A first successful approach would certainly exploits the work of [Fromont et al., 2011] which studies an almost identical question in a simpler case studied.

Chapitre 4

Non reversible optimisation algorithms

In this chapter, I will detail my works derived from the following dynamical system

$$\dot{x}_t = -\frac{1}{t}\int_0^t \nabla U(x_u)du,$$

where U is a real potential defined on $\mathbb{R}^d$ and cohercive for large value of x : $\lim_{|x|\mapsto+\infty} U(x) = +\infty$. Remind firs that such differential equation comes from a numerical modification described in paragraph 1.4 where one aims to find minimum of U. Without loss of generality, we will assume that $\min_{\mathbb{R}^d} U > 0$, and U is at least $\mathcal{C}^2(\mathbb{R}^d)$ and convex sufficiently far from $O_{\mathbb{R}^d}$ [1] :

$$\liminf_{x\to\infty}\langle x, \nabla U(x)\rangle > 0.$$

Moreover, the minimum of U is assumed to be located at point 0 : $U(0) = \min U$.

4.1 Gradient descent with memory model

4.1.1 Physical interpretation

Let be given h and k two smooth non negative and increasing functions, we consider the ordinary differential equation on $\mathbb{R}^d$

$$\dot{x}_t = -\frac{1}{k(t)}\int_0^t h(u)\nabla U(x_u)du. \tag{4.1}$$

A particular natural case will be $k \sim \int h$ for large times t. Using a simple change of variables $t \mapsto \tau(t)$, it is possible to convert (4.1) to a second order differential equation.

Proposition 4.1.1 *Let τ solution of $\dot{\tau} = \sqrt{k(\tau)/h(\tau)}$ and x solution of the memory gradient descent (4.1), then $z = x \circ \tau$ is solution of*

$$\ddot{z}(s) + \gamma(s)\dot{z}(s) + \nabla U(z(s)) = 0, \tag{4.2}$$

where a is a damping function given by $\gamma(s) = \left(\frac{\dot{k}h + k\dot{h}}{2h^{3/2}k^{1/2}}\right) \circ \tau(s)$.

1. We will describe such condition as a mean-reverting property for the differential equation (4.1)

Function γ describes the amount of damping in a dynamical system of an heavy ball which is rolling on a graph of potential U, and submitted to a friction additional force. For some special case of functions h and k, we recover some particular case of the so-called dynamical system *Heavy Ball with Friction* and (4.1) is thus its natural generalization.

From this short physical description, we can expect the dynamical system (4.2) to reach stable critical points of U and maybe the amount of inertia of the ball enables the trajectory to cross some local maxima, which is impossible for standard gradient descent.

First, it is possible to assert the stability of trajectories of o.d.e. (4.2) using quite standard Lyapunov function $\mathcal{E}$ described below as soon as U satisfies an convex-type at infinity condition.

4.1.2 Behaviour of the dynamical system (4.2), convex case

We assume that U satisfies a generalization of convexity described by the following condition [2] :

$$(H_U^1) : \exists \theta > 0 \quad \forall x \in \mathbb{R}^d \qquad U(x) - U(0) \leqslant \theta \langle \nabla U(x), x \rangle.$$

Influence of damping γ The damping effect of γ is important and should be understood as follows : if γ decreases fast to 0, the trajectory (4.2) then possesses an infinite number of oscillations that cannot be insignificant since the dynamical system seems to be almost described by $\ddot{x} + \omega^2 \nabla U x = 0$. This phenomenon is described by the next result.

Proposition 4.1.2 *Let us denote $\mathcal{E}(t) = U(x(t)) + \frac{\dot{x}(t)^2}{2}$, then $\dot{\mathcal{E}}(t) = -\gamma(t)|\dot{x}(t)|^2$ and solutions of (4.2) are defined and bounded on $\mathbb{R}_+$. Moreover,*

$$\forall t > 0 \qquad \mathcal{E}(t) - \min U \geqslant (\mathcal{E}(0) - \min(U)) e^{-\int_0^t \gamma(s)ds}.$$

Hence, even if U is convex, if $\gamma \in \mathbb{L}^1(\mathbb{R}_+)$ the trajectory cannot converge.

The last proposition yields us consider some damping effect $\gamma \notin \mathbb{L}^1(\mathbb{R}_+)$ and it is possible to give a sufficient condition for convergence of $(U(x_t))_{t \geqslant 0}$.

Proposition 4.1.3 *Assume that (H_U^1) is true.*

i) *If γ is a smooth C^1 and non increasing function, then*

$$\int_0^{+\infty} \gamma(s)[\mathcal{E}(s) - \min U]ds < +\infty.$$

ii) *Moreover, if $\int_0^{+\infty} \gamma(s)ds = +\infty$ (slow vanishing damping case), then $\lim \mathcal{E}(t) = \min U$.*

iii) *If there exists $m > 0$ such that $\gamma(t) \geqslant m/t$ for t large enough, then*

$$\mathcal{E}(s) - \min(U) = o\left(\frac{1}{ta(t)}\right).$$

iv) *At last, assume $\arg\min(U) = \{0\}$, then the trajectory converges.*

Remark that this last property does not provide some convergence result of the trajectory itself, this might not be the case when $\arg\min U$ is an open set of $\mathbb{R}^d$. In such situation, it is possible to provide an almost minimal hypothesis.

2. This condition holds as soon as U is convex with $\theta = 1$ for instance.

Theorem 4.1.1 *Assume that* $d = 1$ *and* U *satisfies* (H_U^1) *with* $[\alpha, \beta] \subset \arg\min U$. *If* γ *is such that*

$$\exists k < 1 \qquad \int_0^{+\infty} e^{-k\int_0^s \gamma(u)du} ds < \infty,$$

then the trajectory solution of (4.2) converges. Oppositely, if γ *satisfies*

$$\int_0^{+\infty} e^{-\int_0^s \gamma(u)du} ds = \infty,$$

the trajectory does not converge, except in trivial cases of initialisation in $\arg\min(U)$.

It is also possible to show a similar result for larger dimensions, such details are omitted here owing to their rather technical prerequisite described in [13]. The key assumption to assert convergence or not of the trajectory is the condition $\int_0^{+\infty} e^{-\int_0^s \gamma(u)du} ds = \infty$. At last, we should remark from a technical point of view that the proof depends on a Lyapunov function which enables to control both position and speed of the trajectory, such control is not possible with function $\mathcal{E}$ as pointed by Proposition 4.1.2). This new Lyapunov function uses $\mathcal{E}$ and some additional crossed term between position and speed. Such use of crossing term is quite classical when one considers dissipative system (see some results in [Haraux, 1991] for instance). These additional terms have also been intensively used in works on hypocoercive P.D.E. or probabilistic models.

4.1.3 Behaviour of the dynamical system (4.2), non convex case

We address the typical generic situation where U satisfies the following assumption.
$(\tilde{H}_U)$: U possesses a finite number of m critical points such that $U(x_1) < U(x_2) \cdots < U(x_m)$

Under this last hypothesis, when $\gamma \notin \mathbb{L}^1(\mathbb{R}_+)$, it is possible to show a convergence result on the trajectory (which is slightly weaker than a classical convergence) for the multi-dimensional case.

Theorem 4.1.2 *Assume that* $(\tilde{H}_U)$ *holds, then there exists a unique* x_i *such that*

$$\forall \varepsilon > 0 \qquad \lim_{T\to+\infty} \frac{1}{T} |\{t \leqslant T \quad |x(t) - x_i| > \varepsilon\}| = 0.$$

This ergodic result may be written as

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T x(t)dt = x_i.$$

For the very special case of dimension 1, it is possible to reach a stronger result but its proof is very specific to the one dimensional case (it consists in considering the lengths of time invervals where $|\dot{x}(t)| > 0$).

Theorem 4.1.3 *Assume that all critical points of* U *are non degenerated, i.e.* $U''(x_i) \neq 0$ *and suppose that* γ *is lower bounded as follows* $\gamma(t) \geqslant \frac{c}{1+t}$ *où* $c > 0$. *Then*

i) *For any initialisation point, solutions of (4.2) satisfy* $\lim_{t\to\infty} x(t) = x^\star$ *exists and belongs to* $\{x_1, \dots x_m\}$.

ii) *If* $\mathcal{T}$ *denotes the set of time changes of* $\dot{x}$, *then*

$$|\mathcal{T}| = +\infty \Longleftrightarrow x^\star \textit{is a local minimum of } U.$$

iii) *The set of initialisation points such that* $x^\star$ *is a local minimum is open and dense in* $\mathbb{R}$.

Hence, we do not have obtained satisfactory result on the convergence towards the global minimum of U, it was thus quite natural to be interested by some noisy perturbation of the dynamical system (4.1).

4.2 Memory average gradient diffusion

4.2.1 Average diffusion model

We describe in this paragraph a natural generalization of (4.1) when the dynamical system is corrupted by a standard Brownian motion. We still consider two increasing maps h and k which are non negatives. If σ is an invertible squared covariance matrix of size d and $(B_t)_{t\geqslant 0}$ a d-dimensional Brownian motion, the dynamical system is then described by the following stochastic differential equation

$$dX_t = -\frac{1}{k(t)}\left(\int_0^t h(u)\nabla U(X_u)du\right)dt + \sigma dW_t, \tag{4.3}$$

We define $(Y_t)_{t\geqslant 0}$ the instantaneous drift of (X_t),

$$Y_t = \frac{1}{k(t)}\int_0^t h(s)\nabla U(X_s)ds,$$

and we remark that $dY_t = (h/k)(t)(\nabla U(X_t) - Y_t)dt$. Hence, (4.3) is a kinetic differential system $2d$-dimensional, it is also an inhomogeneou Markov process described by :

$$\begin{cases} dX_t = \sigma(X_t)dW_t - Y_t dt. \\ dY_t = r(t)(\nabla U(X_t) - Y_t)dt, \end{cases} \tag{4.4}$$

where $r(t) = \frac{h}{k}(t)$ is $\mathcal{C}^1$.

In the sequel, I will only discuss on the case $h = \dot{k}$ even if it is also possible to extend some of these results to more general memory cases. It is easy to check that $(X_t, Y_t, t)_{t\geqslant 0}$ is an homogeneous Markov process whose generator $\mathcal{A}$ acts on $f \in \mathcal{C}^2_K(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}_+)$ following :

$$\mathcal{A}f(x,y,t) = -\langle y, \nabla_x f\rangle + r(t)\langle \nabla U(x) - y, \nabla_y f\rangle + \frac{1}{2}\mathrm{Tr}\left(\sigma^*(x)D_x^2 f(x,y)\sigma(x)\right) + \partial_t f. \tag{4.5}$$

We will assume that U satisfies the assumption (H_U) given by :

Assumption 10 ($\mathbf{H_U}$) $\lim_{|x|\to+\infty} U(x) = +\infty \quad \liminf_{|x|\to+\infty}\langle x, \nabla U(x)\rangle > 0, \quad Tr\left[\sigma^* D^2 U \sigma\right] \leqslant CU.$

This assumption is true for a wide class of potentials U : for instance $U(x) \sim_\infty C_1|x|^p$ with $D^2U(x) \sim_\infty C_2|x|^{p-2}$ satifies $(\mathbf{H_U})$ as soon as $\|\sigma(x)\| = O(|x|)$. This is also the case for weaker increasing U : $U(x) \sim_\infty C_1 \ln|x|$ and $D^2U(x) \sim_\infty C_2|x|^{-2}$ with $\|\sigma(x)\| = O(1+|x|)$ also satisfies $(\mathbf{H_U})$.

Proposition 4.2.1 *Assume that $(\mathbf{H_U})$ holds, then there exists a unique strong solution of (4.4). Moreover, if (X_0, Y_0) is such that $\mathbb{E}[U(X_0) + |Y_0|^2] < +\infty$, then for any $T > 0$*

$$\sup_{t\in[0,T]} \mathbb{E}[U(X_t) + |Y_t|^2] < +\infty.$$

The proof relies on a control within a finite time of trajectories and uses a Gronwall Lemma on the classical Lyapunov function defined as :

$$\mathcal{E}(x,y,t) = U(x) + \frac{|y|^2}{2r(t)}. \tag{4.6}$$

We will study in the sequel the regularity of the semi-group associated to $(X_t, Y_t, t)_{t\geqslant 0}$ as well as the convergence to steady regime (when one steady regime exists). For sake of simplicity, we will note $z_0 = (x_0, y_0) \in \mathbb{R}^d \times \mathbb{R}^d$ the initialisation point (random or not) of the diffusion.

4.2.2 Hypo-ellipticity

The random process (4.4) is totally degenerated on coordinate Y, thus existence of density and regularity properties of $P_t(z_0, .)$ is not so clear. We next provide two important results in order to obtain irreducibility of the Markov process [3].

Existence and regularity of the density with respect to the Lebesgue measure The first result concerns the existence of density with respect to the Lebesgue measure and uses the set $\mathcal{E}_U$ defined as

$$\mathcal{E}_U = \left\{x \in \mathbb{R}^d,\ \det\left(D^2U(x)\right) \neq 0\right\}, \qquad \text{et} \qquad \mathcal{M}_U = \mathbb{R}^d \backslash \mathcal{E}_U. \tag{4.7}$$

We then assume that :

Assumption 11 ($\mathbf{H_{Hypo}}$) *σ and U are $\mathcal{C}^\infty$ and there exists $\varepsilon_0 > 0$ such that $\sigma\sigma^* \geqslant \varepsilon_0 \mathrm{Id}$, (uniform ellipticity of σ over $\mathbb{R}^d$). Moreover, the manifold $\mathcal{M}_U$ is such that $\dim(\mathcal{M}_U) \leqslant d-1$.*

The vector fields that correspond to the diffusion part and the drift part in (4.4) are

$$L_\sigma(x)(f) = \frac{1}{2}\sum_{j=1}^{d} \langle \nabla_x(\sigma_j)(x), \sigma_j(x)(f)\rangle.$$

where

$$\forall j \in \{1 \dots d\} : \qquad \sigma_j(x) = \sum_{i=1}^{d} \sigma_j^i(x)\partial_{x_i}. \tag{4.8}$$

and

$$L_D(t,x,y) = -\langle y, \nabla_x\rangle + r(t)\langle \nabla U(x) - y, \nabla_y\rangle.$$

Proposition 4.2.2 *Assume that ($\mathbf{H_{Hypo}}$) holds, then for any $z_0 \in \mathbb{R}^d \times \mathbb{R}^d$ and any $t > 0$, $P_t(z_0, .)$ is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^d \times \mathbb{R}^d$. Moreover, for any $t > 0$ and $z_0 \in \mathbb{R}^d \times \mathbb{R}^d$, $z \mapsto p_t(z_0, z)$ is $\mathcal{C}^\infty$ over $\mathbb{R}^d \times \mathbb{R}^d$ where $p_t(z_0, .)$ is the density of $P_t(z_0, .)$.*

This proposition uses the fact that the dimension of the Lie algebra spanned by $\partial_t + (L_D - L_\sigma), \sigma_1, \dots, \sigma_d$ is $2d + 1$ under assumption ($\mathbf{H_{Hypo}}$), it is thus possible to use Hörmander theorem. This proposition does not give any result on the smoothness of $(z_0, z) \mapsto p_t(z_0, z)$, this application should certainly be continuous if one assume moreover that vector fields possess at the most polynomial growth and this question could be tackled using Malliavin calculus or Harnack inequality (see for instance[Hairer, 2011] or [Pascucci and Polidoro, 2006]) but this point has been get rounded and still remains open since I have note studied this question.

3. Irreducibility is especially important when the process is homogeneous in order to obtain uniqueness of invariant measures.

Minoration of $p_t(z_0,.)$ Positiveness of $p_t(z_0,.)$ (defined in the paragraph above) is indeed rather different from the use Hörmander condition to obtain smooth density w.r.t. Lebesgue measure. In fact, a minoration of $p_t(z_0,z)$ traduces that a sufficient amount of trajectories of (4.4) starting at point z_0 can reach neighbourhoods of z. It is thus a problem of control for trajectories defined by the following differential system (4.9).

$$\begin{cases} \dot{x}(t) = \sigma(x(t))\varphi(t) - y(t). \\ \dot{y}(t) = r(t)(\nabla U(x(t)) - y(t))dt, \end{cases} \tag{4.9}$$

Controllability of such differential system will be discussed in detail in the last section of this chapter, but this controllability is already important here. It is clear that starting from any z_0 of $\mathbb{R}^d \times \mathbb{R}^d$, one can reach any arbitrary point on coordinate x, but this is largely more complicated on the coordinate y since the control φ only acts on $x(t)$ and not on $y(t)$. Indeed, we should consider the initial problem and remark that $y(t)$ can still be written as

$$y(t) = y_0 + \frac{1}{k(t)} \int_0^t \dot{k}(s)\nabla U(x(s)). \tag{4.10}$$

Hence, if ∇U is bounded, y cannot exit from $B(y_0, \|\nabla U\|_\infty)$ and equation (4.10) naturally stimulates us to assume that ∇U is surjective in$\mathbb{R}^d$. Following such idea, we then obtain the following result.

Proposition 4.2.3 *Assume that* $(\mathbf{H_{Hypo}})$ *is true and that* $\lim_{|x|\to+\infty} \frac{U(x)}{|x|} = +\infty$, *then the two following points are satisfied.*
(i) For any $T > 0$ *and* $z_0 \in \mathbb{R}^d \times \mathbb{R}^d$, *if* $\mathcal{O} \subset \mathbb{R}^d \times \mathbb{R}^d$ *is an arbitrary open set, then* $P_T(z_0, \mathcal{O}) > 0$. *Hence, for all*$z_0 \in \mathbb{R}^{2d}$, $p_T(z_0,.)$ *is* λ_{2d}*- a.s. positive and there exists at the most a unique invariant measure for* $(X_t, Y_t)_{t\geqslant 0}$ *when* $r(t) \longmapsto r_\infty \in (0; +\infty)$.
(ii) Assume r *to be a positive constant and that there exists a minimum* x^* *of* U *such that* $D^2U(x^*)$ *is invertible, then if we denote* $z^* = (x^*, 0)$, *one can find* $T > 0$ *such that for any compact* K *of* $\mathbb{R}^{2d}$, *one can find* $\nu_K > 0$ *and* $\alpha(T, K) > 0$ *such that*

$$\forall z_0 \in K, \quad P_T(z_0,.) \geqslant \alpha(T,K)\lambda_{2d}(. \cap B(z^*, \nu_K)).$$

The first point uses the controllability of the differential system (4.9) and the Fenchel-Legendre transform of U : if $\lim_{|x|\to+\infty} \frac{U(x)}{|x|} = +\infty$, for any ν in $\mathbb{R}^d$, the map $F_\nu(x) = \langle \nu, x\rangle - U(x)$ has a maximumand thus ∇U is surjective. In order to reach any open set $\mathcal{O}$ of $\mathbb{R}^d \times \mathbb{R}^d$, we then build a trajectory in three parts : the first one bring x to $x(\eta)$, the second part remains constant in coordinate x between η and $T-\eta$ then the last part bring the trajectory into $\Pi_x(\mathcal{O})$. Of course, we must find $x(\eta)$ so that the time spent between η and $T-\eta$ on this point enables $y(T-\eta)$ to reach $\Pi_y(\mathcal{O})$. Such point $x(\eta)$ exists owing to the surjectivity of ∇U. This strategy permits to show the approached controllability of the differential system (4.9).

The second point is crucial in order to show that compact sets are *petite sets* for the application of Meyn and Tweedie estimates. The proof relies on a stronger result on the controlled system (4.9) which stands for the exact small controllability around the equilibrium z^*. This is ensured by the non degeneracy of $D^2U(x^*)$ which implies the full rank Kalman condition for the linearised system around z^* (one may find further details in[Coron, 2007] for instance). This last condition should be replaced by any other sufficient conditions which implies the exact local controllability near z^*. This controllability result is then sufficient to obtain enough mass around z^* to obtain lower bound on p_t following the argument of[Delarue and Menozzi, 2010].

4.2.3 Steady regimes ($r_\infty > 0$)

Short range memory We describe in this part some results on the asymptotic behaviour of (X_t, Y_t) when the memory (described by application $t \mapsto r(t)$) is not too long. Such situation is identified through the limit behaviour of r at $+\infty$. Such steady regime corresponds to $r(t) \longmapsto r_\infty \in]0, +\infty]$. We thus assume the following hypothesis on r.

Assumption 12 ($\mathbf{H_r}$) *The map r possesses a non negative limit r_∞ when $t \longmapsto +\infty$ (where $r_\infty = +\infty$ is admissible). Moreover, we assume that r varies slow enough near $+\infty$:*

$$\lim_{t\mapsto+\infty} \frac{r'(t)}{r^2(t)} = 0.$$

This last assumption is encountered in the two following situations :

- $k(t) = \exp(\lambda t)$ and $r(t) = r_\infty = \lambda$ thus $(X_t^z, Y_t^z)_{t\geqslant 0}$ is homogeneous Markov.
- $k(t) = \exp(t^\alpha)$ with $\alpha > 1$ and in this case $r_\infty = \lim_{t\to+\infty} r(t) = +\infty$.

Lyapunov function The stability of the process is guaranteed as soon as ∇U possesses enough repelling force to imply a tightness property. This is traduced by the following somewhat technical hypothesis.

Assumption 13 ($\tilde{\mathbf{H}}_\mathbf{U}$) *There exists $m \in (0, r_\infty)$ and $\varepsilon \in (0, r_\infty - m)$ such that*

$$\limsup_{|x|\to+\infty} \left(-m\langle x, \nabla U(x)\rangle + \frac{1}{2}\mathrm{Tr}\left(\sigma^*(x)(D^2U(x) + (m+\varepsilon)I_d)\sigma(x)\right)\right) = -\infty.$$

$(\tilde{\mathbf{H}}_\mathbf{U})$ is stronger thatn assumption $(\tilde{\mathbf{H}}_\mathbf{U})$ but is not too restrictive. If σ is independent from x, $(\tilde{\mathbf{H}}_\mathbf{U})$ holds for potentials $U(x) \sim_{|x|\to+\infty} |x|^q$ as soon as $q > 0$ and it is even true when $U(x) \sim_{|x|\mapsto+\infty} \ln(|x|+1)^\beta$ with $\beta > 1$. σ may also vary, but it should not be too large when $x \mapsto \infty$:

- For polynomial growth of U : $U(x) \sim_{|x|\to+\infty} |x|^q$ with $q > 0$, this assumption is true for $\|\sigma(x)\sigma^*(x)\| = o(|x|^{q\wedge 2})$ and $|x| \to +\infty$.
- For logarithmic growth of U : $U(x) \sim_{|x|\to+\infty} \ln(|x|+1)^\beta$ with $\beta > 1$, this assumption is true if $\|\sigma(x)\sigma^*(x)\| = o(\ln(|x|+1)^{\beta-1})$ when $|x| \to +\infty$.

The key point is now to build a Lyapunov function which permits to control the dynamical system both on coordinate x and y. One should remark that the former classical function $\mathcal{E}$ defined by $\mathcal{E}(x, y, t) = U(x) + \frac{|y|^2}{2r(t)}$ does not satisfy such requirements since only coordinate y is bounded :

$$\mathcal{A}\mathcal{E}(x, y, t) = -y^2\left(1 + \frac{r'(t)}{2r^2(t)}\right) + \frac{1}{2}\mathrm{Tr}\left(\sigma^*(x)D^2U(x)\sigma(x)\right).$$

However, it is possible to slightly modify this function by the addition of a crossed term « position - speed »to obtain also information on coordinate x. For a given couple $(m_\varepsilon, \varepsilon)$ described in $(\tilde{\mathbf{H}}_\mathbf{U})$, we define

$$V(x, y, t) = U(x) + \frac{|y|^2}{2r(t)} + m_\varepsilon\left(\frac{|x|^2}{2} - \frac{\langle x, y\rangle}{\rho(t)}\right), \tag{4.11}$$

where ρ is a real function solution of the o.d.e.

$$\rho(t) = \left(\int_t^{+\infty} \frac{k(t)}{k(s)}ds\right)^{-1}.$$

Function V describes a real repelling force on coordinates x and y since for t large enough

$$\mathcal{A}V(x, y, t) \leqslant -C_1\langle x, \nabla U(x)\rangle + \frac{1}{2}\mathrm{Tr}\left(\sigma^*(x)D^2U(x)\sigma(x)\right) - C_2|y|^2.$$

Occupation measures For $z_0 \in \mathbb{R}^d \times \mathbb{R}^d$, we consider the two family of occupation measures $(\nu_t^{z_0}(\omega, dx, dy))_{t\geqslant 1}$ and $(\mu_t^{z_0}(dx, dy))_{t\geqslant 1}$ defined as follows : for any bounded measurable continuous $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we denote :

$$\nu_t^{z_0}(\omega, f) = \frac{1}{t}\int_0^t f(X_s^{z_0}, Y_s^{z_0})ds,$$

and

$$\mu_t^{z_0}(f) = \frac{1}{t}\int_0^t \mathbb{E}[f(X_s^{z_0}, Y_s^{z_0})]ds = \mathbb{E}[\nu_t^{z_0}(\omega, f)].$$

It is possible to show ergodicity of $(\mu_t^{z_0})_{t\geqslant 0}$:

Theorem 4.2.1 *Assume that* $(\tilde{\mathbf{H}}_U)$ *and* $(\mathbf{H_r})$ *hold with* $r_\infty \in \mathbb{R}_+^* \cup \{+\infty\}$, *for any* $z_0 =\in \mathbb{R}^d \times \mathbb{R}^d$, $(\mu_t^{z_0})_{t\geqslant 1}$ *is tight. If* μ_∞ *denotes any accumulation point of* $(\mu_t^z)_{t\geqslant 1}$ *when* $t \to +\infty$, *one has*

(i) If $r_\infty = +\infty$, *the first marginal (on coordinate* x*) of* μ_∞ *is an invariant measure of the Kolmogorov process*

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

(ii) If $r(t) \xrightarrow{t\to+\infty} r_\infty < +\infty$, μ_∞ *is an invariant distribution of the homogeneous Markov process solution of* (4.4) *with* $r(t) = r_\infty$, $\forall t \geqslant 0$.

It is also possible to obtain a convergence result on the random occupation measures $(\nu_t^{z_0}(\omega, dx, dy))_{t\geqslant 1}$ under the following slightly stronger hypothesis.

Assumption 14 $(\check{\mathbf{H}}_U)$ *There exists* $a \in (0, 1]$, $\beta \in \mathbb{R}$ *and* $\alpha > 0$ *such that*

(i) $\quad -\langle x, \nabla U(x)\rangle \leqslant \beta - \alpha\left(U(x) \vee |x|^2\right)^a, \forall x \in \mathbb{R}^d$

(ii) $\quad (1 + \mathrm{Tr}(\sigma\sigma^*)(x))\left(1 + \frac{|\nabla U(x)|^2}{U(x)} + \|D^2U(x)\| + |||D^3U(x)|||\right) \overset{|x|\to+\infty}{=} o((U(x) \vee |x|^2)^a).$

Under such condition, it is possible to show similar results on $(\nu_t^{z_0}(\omega, dx, dy))_{t\geqslant 1}$ which are described in [15].

Stationary measures and convergence rates It is possible to describe the nature of the equilibrium of the process (X_t, Y_t) in the case$0 < r_\infty < +\infty$.

Proposition 4.2.4 *Assume that* $(\mathbf{H_r})$, $(\mathbf{H_{Hypo}})$ *and* $(\tilde{\mathbf{H}}_U)$ *are true and suppose that* $r(t) = r_\infty \in \mathbb{R}_+^*$. *If* $\lim_{|x|\to+\infty} \frac{U(x)}{|x|} = +\infty$, *then there exists a unique invariant measure* ν *such that*

- *(i)* ν *is absolutely continuous w.r.t. the Lebesgue measure, with density* p_{r_∞} *which is* $\mathcal{C}^\infty(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}_+)$. *Moreover,* p_{r_∞} *is the unique probability measure solution of*

$$\langle y, \nabla_x p_{r_\infty}\rangle + \frac{1}{2}\mathrm{Tr}\left(\sigma^* D_x^2 p_{r_\infty}\sigma)\right) + r_\infty\left[\langle y - \nabla U(x), \nabla_y p_{r_\infty}\rangle + p_{r_\infty}\right] = 0. \tag{4.12}$$

- *(ii) If* $d = 1$, $U(x) = x^2/2$ *and* $\sigma(x) = \sigma > 0$ $\forall x \in \mathbb{R}$, *and* $r(t) = r_\infty \in]0; +\infty[$, *then* p_{r_∞} *is a Gaussian measure centered with covariance matrix*

$$\Sigma^2(r_\infty) = \frac{\sigma^2}{2}\begin{pmatrix} \frac{r_\infty+1}{r_\infty} & 1 \\ 1 & 1 \end{pmatrix}.$$

Remark 4.2.1 *The situation is quite simple when* $r_\infty = +\infty$ *since Theorem 4.2.1 shows that the limiting behaviour of the memory diffusion is similar to the Kolmogorov one, but when* $r_\infty \in (0, +\infty)$ *the limiting invariant measure is non standard since even in the Gaussian case, the density* p_{r_∞} *is a twisted Gaussian. The more* r_∞ *is near* 0*, the more longer is the memory which leads to an explosive variance of* p_{r_∞}*. In the general case, the P.D.E. satisfied by* p_{r_∞} *does not seem to possess explicit solutions.*

At last, it is possible to describe convergence rate results of $P_t(z_0, .)$ when $t \to +\infty$.
In the homogeneous case $r(t) = r_\infty$, it is possible to use the approach of [Down et al., 1995].

Theorem 4.2.2 *Assume that* r *is homogeneous :* $r(t) = r_\infty > 0$ *and that hypo-elliptic assumptions of proposition 4.2.3, ii) are true. If* U *satisfies (*$\check{\mathbf{H}}_U$*) for some* $a \in (0, 1]$*, then for any* $p \geqslant 1$ *and* $t \geqslant 0$ *:*

$$\sup_{\{f,|f|\leqslant 1\}} |P_t^{r_\infty}(z_0, f) - \nu(f)| \leqslant C_{a,p,r_\infty} V_\infty^p(z_0) \begin{cases} \exp(-\gamma_{p,r_\infty} t) & \text{if } a = 1 \\ t^{-\frac{p+a-1}{1-a}} & \text{if } a \in (0,1). \end{cases}$$

where $z = (x, y)$*,* V_∞ *is a positive function defined as* $V_\infty(z) = U(x) + \frac{r_\infty}{2}\left|x - \frac{y}{r_\infty}\right|^2$*,* γ_{p,r_∞} *and* C_{a,p,r_∞} *are explicit non negarive constants that does not depend on* z_0 *and* t*.*

Remark that it is also possible to give some convergence rates when $r_\infty = +\infty$ using some coupling argument to the diffusion $dX_t = -\nabla U(X_t)dt + \sigma(X_t)dB_t$. Further details may be found in [15].

4.2.4 Explosion ($r_\infty = 0$)

When the memory function r satisfies $\lim_{t\mapsto+\infty} r(t) = r_\infty = 0$, we have a long memory in the process. Such typical case is $k(t) = (1 + t)^\alpha$ for any $\alpha > 0$ or when $k(t) = e^{(1+t)^\alpha}$ with $0 < \alpha < 1$.

Under-quadratic potential We have obtained a quite precise result in the under-quadratic case of potential U. This result is summarized in the following theorem.

Theorem 4.2.3 *Assume that there exists* C *such that* $|\nabla U|^2 \leqslant C(1 + U)$ *and* $\lambda_0 > 0$ *for which* $\mathrm{Tr}(\sigma^* D^2 U \sigma)(x) \geqslant \lambda_0 > 0$*. If* $r_\infty = 0$ *and for* t *large enough* $r'(t) + 2r^2(t) \geqslant 0$*, then for any* z_0

$$\limsup_{t\to+\infty} r(t)\mathbb{E}[|X_t^{z_0}|^2] > 0.$$

Moreover, there exists a sequence $(t_n)_{n\geqslant 1}$ *such that* $\mathbb{E}[|X_{t_n}^{z_0}|^2] \to +\infty$*.*

This theorem may be applied for instance when the weighting memory is uniform all along the trajectory before t : $Y_t = \frac{1}{1+t}\int_0^t \nabla U(X_s)ds$. Hence, to obtain a stable process with a long range memory, it is necessary to reduce the volatility of the random dynamical system. This phenomenon may be explained by the analogy with the phisical interpretation of the HBF model. More details are given in the introduction of [15].

Quadratic potential It is also possible to obtain a very precise behaviour when U is quadratic. If one considers the result of proposition 4.2.4 ii), we may remark that when $r_\infty \mapsto 0$, the covariance matrix becomes « infinite »on coordinate x. Since $(X_t, Y_t)_{t\geqslant 0}$ is a Gaussian process, if there exists an invariant measure, this latter one should also be Gaussian, which yields finally a non existence result.

Moreover, we assume $U(x) = x^2/2$, $d = 1$ and that the memory is polynomial : $k(t) = (1+t)^\alpha$ (thus $r(t) = \alpha/(1+t)$). All information is given in $f(t) = \mathbb{E}[X_t^2]$, $g(t) = \mathbb{E}[Y_t^2]$ and $h(t) = \mathbb{E}[X_t Y_t]$. Itô's formula shows that

$$(\mathcal{S}) \begin{cases} f'(t) = & 1 - 2h(t) \\ g'(t) = & 2r(t)[h(t) - g(t)] \\ h'(t) = & -g(t) + r(t)[f(t) - h(t)]. \end{cases}$$

We then obtain the following theorem.

Theorem 4.2.4 *Assume* $d = 1$, $U(x) = x^2/2$ *and* $k(t) = (1+t)^\alpha$ *with* $\alpha > 1/2$, *one has :*

i) For any z_0, $(X_t^{z_0}, Y_t^{z_0})_{t\geqslant 0}$ *is asymptotically centered.*

ii) The process $(X_t^{z_0}, Y_t^{z_0})_{t\geqslant 0}$ *satisfies*

$$\lim_{t\to\infty} \mathbb{E}Y_t^2 = \frac{\alpha}{2\alpha+1}, \qquad \textit{and} \qquad \mathbb{E}X_t^2 \sim \frac{t}{2\alpha+1} \quad \textit{when} \quad t \to +\infty.$$

iii)

$$\left(\sqrt{\frac{2\alpha+1}{t}} X_t, \sqrt{\frac{2\alpha+1}{\alpha}} Y_t\right) \overset{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0, I_2) \quad \textit{when} \quad t \to +\infty.$$

4.3 Particular case of kinetic Fokker-Planck evolutions

4.3.1 Model

When we have thought about average diffusion (4.4), we were mainly interested in properties that may naturally be used for optimization applications. Thus, the main questions of first interest for such applications are behaviour of invariant measures with small parameters, and the nature of the evolution of the norm $\|P_t - \mu_\infty\|_{\mathbb{L}^2(\mu_\infty)\circlearrowleft}$ when t goes to $+\infty$. It is important to well estimate the constants which describe the exponential convergences to equilibrium in order to obtain the best simulated annealing as possible. Results described in [15] are in some sense quite unsatisfactory since we only have results in total variation norm and constants may not be so good for such particular kinetic equation. In [19], we study a situation which is more known : the kinetic Fokker-Planck evolution. Such equations are defined through the semi-group $(P_t)_{t\geqslant 0}$ givent by

$$\begin{cases} dX_t = \sigma(V_t)dt. \\ dV_t = -\nabla U(V_t) + a dW_t, \end{cases} \tag{4.13}$$

where a is a non negative parameter and W_t a standard Brownian motion. Even if there does not exists a linear change of variables which permits to deduce results on (4.4) from (4.13) [4], these equations are quite similar, at least from a visual point of view. It was thus natural to consider the computation of the norms $\|P_t - \mu_\infty\|_{\mathbb{L}^2(\mu_a)\circlearrowleft}$ where μ_a is the stationary measure of Fokker-Planck kinetic semi-group(4.13) which is explicit here (contrary to the one of (4.4)).

4. In the Gaussian case, one can write (4.4) using a different parametrisation to obtain $dX_t = Y_t dt$ and $dY_t = -(X_t + Y_t)dt + dW_t$

4.3.2 Norm computation $\mathbb{L}^2(\mu_a)_{\circlearrowleft}$ for $U = 0$.

In [19], we provide some exact results for the norm $\mathbb{L}^2(\mu_a)_{\circlearrowleft}$ when considering the Fokker-Planck kinetic semi-group in very particular cases.

The first toy model is reduced to the state space $\mathbb{T} \times \mathbb{R}$ for the "position $\times$ speed" where $\mathbb{T} := \mathbb{R}/(2\pi\mathbb{Z})$. Let be given any $a > 0$, the operator of interest is

$$L_a = y\partial_x + a\partial_y^2 - y\partial_y, \tag{4.14}$$

which is a particular case of(4.13) when $U = 0$ and the position lives in a compact set. It is quite easy to see that $P_t^{(a)}$ converges towards $\mu_a = \lambda \otimes \gamma_a$ where λ is the uniform law on $\mathbb{T}$ and $gamma_a$ is the Gaussian distribution centered with variance a. After some quite technical developments, it is possible to compute the evolution of the norm.

Theorem 4.3.1 *For any $a > 0$ and $t \geqslant 0$, we have*

$$\|P_t^{(a)} - \mu_a\|_{\mathbb{L}^2(\mu_a)_{\circlearrowleft}} = \max\left(\exp(-t), \exp\left[-a\left(t - 2\frac{1-\exp(-t)}{1+\exp(-t)}\right)\right]\right), \tag{4.15}$$

where $\|.\|_{\mathbb{L}^2(\mu_a)_{\circlearrowleft}}$ is the operator norm in $\mathbb{L}^2(\mu_a)$.

The proof relies on a natural decomposition of the generator L_a on a basis of $\mathbb{L}^2(\mu_a)$ obtained by a tensor product of trigonometric polynomials in coordinate x and Hermite polynomials in coordinate y. We then identify some infinite dimensional orthogonal subspaces which are stable by L_a, denoted $\mathcal{V}_{p_{p\geqslant 0}}$ in this formal description. L_a act on each $\mathcal{V}_{p_{p\geqslant 0}}$ as an infinite tri-diagonal anti-symmetric matrix. We should note that this is this anti-symmetry which represents a serious problem for the computation of the norm of L_a since eigenvalues of L_a are no longer orthogonal in $\mathbb{L}^2(\mu_a)$.

The key point which permits to compute both eigenvalues and eigenvectors of L_a on each $\mathcal{V}_p$ is to decompose the operator in $D + c_{a,p}S - c_{a,p}S^\star$ and then use the Lie algebra spanned by D, S et $S^\star$ which is three dimensional here. This important property enables the complete computation of the spectrum of L_a in this precise case, the eigenvalues are all reals for any value of a, as well as all associated eigenvectors. Instead of more details on the computations which are extremely technicals [5], let us comment on some numerical conclusions brought by Theorem 4.3.1.

4.3.3 Qualitative behaviour, $U = 0$

It is above all interesting to look at the asymptotic behaviour of the norm computed by Theorem4.3.1 for small and large times. When $t \mapsto 0_+$,

$$\ln\left(\|P_t^{(a)} - \mu_a\|_{\mathcal{L}^2(\mu_{a/c})_{\circlearrowleft}}\right) = -\frac{a}{12}t^3(1 + o(1)). \tag{4.16}$$

This shows that the norm decreases very slowly at the beginning of the evolution and the power 3 should be considered as the first order of hypo-coercivity of operator L_a. Moreover, when t growth to $+\infty$,

$$-\ln\left(\|P_t^{(a)} - \mu_a\|_{\mathcal{L}^2(\mu_{a/c})_{\circlearrowleft}}\right) = \begin{cases} a(t - 2 + \mathcal{O}(e^{-t})) & \text{, if } a \leqslant 1 \\ t & \text{, if } a > 1, \end{cases}$$

5. Commented as "nearby overkill" by some reader

which stands for the convergence to equilibrium of the semi-group $(P_t^{(a)})_{t\geqslant 0}$. This bound is of course coherent with former bounds obtained in general studies for kinetic Fokker-Planck equations but constants here are explicit. Using a scaling argument, it is possible to deduce from Theorem 4.3.1 the following corollary :

Corollary 4.3.1 *For any* $a, c > 0$ *and* $b \in \mathbb{R}\backslash\{0\}$, *we consider*

$$L_{a,b,c} := by\partial_x + a\partial_y^2 - cy\partial_y \tag{4.17}$$

which admits for invariant measure $\mu_{a/c}$, *then the semi-group* $(P_t^{(a,b,c)})_{t\geqslant 0}$ *satisfies*

$$\forall t \geqslant 0, \qquad \|P_t^{(a,b,c)} - \mu_{a/c}\|_{\mathcal{L}^2(\mu_{a/c})\circlearrowleft} = \max\left(\exp(-ct), \exp\left[-\frac{ab^2}{c^3}\left(ct - 2\frac{1-\exp(-ct)}{1+\exp(-ct)}\right)\right]\right).$$

In particular, the convergence to equilibrium is given by

$$\lim_{t\mapsto+\infty} -\frac{1}{t}\ln\left(\|P_t^{(a,b,c)} - \mu_{a/c}\|_{\mathcal{L}^2(\mu_{a/c})\circlearrowleft}\right) = \min\left(c, \frac{ab^2}{c^2}\right).$$

It is quite tempting to compare this convergence rate to the usual one obtained with the Heat kernel $(Q_t^{(a)})_{t\geqslant 0}$ on $\mathbb{T}$ generated by the operator $K_a := a\partial_x^2$. K_a uses at each time the same amount of randomness as $L_{a,b,c}$ (where $b \in \mathbb{R}$ and $c > 0$ which are tuning parameters). Since K_a is self-adjoint in $\mathbb{L}^2(\lambda)$ and admits a spectral gap of a, we have

$$\forall t \geqslant 0, \qquad \|Q_t^{(a)} - \lambda\|_{\mathbb{L}^2(\lambda)\circlearrowleft} = \exp(-at).$$

Thus, if one considers a Monte Carlo method to simulate the uniform law λ, it would be useful to choose the hypo-coercive simulation $(P_t^{(a,b,c)})_{t\geqslant 0}$ instead of $(Q_t^{(a)})_{t\geqslant 0}$ with the choice $c > a$ and $b/c > 1$ and then project the simulations on the first coordinate. Of course, this is just an example since the simulation of a Brownian motion is clearly more costly than a simple simulation of a uniform law but it shows that equilibrium convergence rates can be improved by the use of non reversible dynamics. Works of [Diaconis et al., 2010b] has already shown such phenomenon in a framework of second order Markov chains.

4.3.4 Hypo-coercive Ornstein-Uhlenbeck process.

We now describe briefly the results on the hypo-coercive Ornstein-Uhlenbeck process defined on $\mathbb{R}\times\mathbb{R}$ through

$$\tilde{L}_a := y\partial_x + -ax\partial_y + \partial_y^2 - y\partial_y. \tag{4.18}$$

The stationary measure is still explicit here and given by $\tilde{\mu}_a := \gamma_{1/a}\otimes\gamma_1$. We are going to study the semi-group evolution $(\tilde{P}_t^{(a)})_{t\geqslant 0}$ in $\mathbb{L}^2(\tilde{\mu}_a)$. The idea is again to compute the effect of $\tilde{L}_a$ on a basis of $\mathbb{L}^2(\tilde{\mu}_a)$ obtained by the tensor product of Hermite polynomials in variables x and y. We first identify orthogonal subspaces which are stable by $\tilde{L}_a$ and let us denote them $\tilde{\mathcal{V}}_p$. On these spaces, $\tilde{L}_a$ may be decomposed in a similar way as it was also the case when $U = 0$, $\tilde{L}_a = \tilde{D} + \tilde{c}_{a,p}\tilde{S} - \tilde{c}_{a,p}\tilde{S}^\star$ on $\tilde{\mathcal{V}}_p$. This enables the exact computation of the eigenvectors and eigenvalues of $\tilde{L}_a$ on each chaque $\tilde{\mathcal{V}}_p$ and next on the whole space $\mathbb{L}^2(\tilde{\mu}_a)$. Here, the spectrum possesses a different behaviour according to the position of a with respect to $1/4$: if $a < 1/4$, the spectrum is real and $\tilde{L}_a$ is diagonalisable in a non orthonormal basis of $\mathbb{L}^2(\tilde{\mu}_a)$. If $a > 1/4$, the same property still holds even if the spectrum is not real. At last, if $a = 1/4$, $\tilde{L}_a$ is no longer diagonalisable and has Jordan blocks of all order. At last, $\tilde{P}_t^{(a)} - \tilde{\mu}_a$ can be computed in $\mathbb{L}^2(\tilde{\mu}_a)$.

Theorem 4.3.2 *For all* $a > 0$ *and* $t \geqslant 0$, *one has*

$$\|\tilde{P}_t^{(a)} - \tilde{\mu}_a\|_{\mathcal{L}^2(\mu_{a/c})\circlearrowleft} = C_a(t) \exp\left(-\frac{1-\sqrt{(1-4a)_+}}{2}t\right), \tag{4.19}$$

where $\|.\|_{\mathcal{L}^2(\mu_{a/c})\circlearrowleft}$ *is the operator norm in* $\mathbb{L}^2(\tilde{\mu}_a)$ *and* $C_a(t)$ *is given by*

– *If* $a \in (0, 1/4)$, *denote* $\theta := \sqrt{1-4a}$ *and*

$$C_a(t) := \sqrt{e^{-\theta t} + \frac{1-\theta^2}{2\theta^2}(1-e^{-\theta t})^2 + \frac{1-e^{-2\theta t}}{2}\left(1 + \frac{1}{\theta}\sqrt{1+(\theta^{-2}-1)\left(\frac{e^{\theta t}-1}{e^{\theta t}+1}\right)^2}\right)}.$$

– *If* $a \in (1/4, +\infty)$, *denote* $\theta := \sqrt{4a-1}i$ *and*

$$C_a(t) := \sqrt{1 + \frac{|e^{\theta t}-1|}{2|\theta|^2}\left(|e^{\theta t}-1| + \sqrt{|e^{\theta t}-1|^2 + 4|\theta|^2}\right)}.$$

– *If* $a = 1/4$,

$$C_a(t) := \sqrt{1 + \frac{t^2}{2} + t\sqrt{1+\left(\frac{t}{2}\right)^2}}.$$

Again, if t is small enough, the decreasing power is three (see [19] for precise computations) although when $t \mapsto +\infty$, we get an exponential convergence whose rate depends on the position of a regarding $1/4$. If $a > 1/4$, the map $C_a(t)$ oscillates with a period $T_a = 2\pi/\sqrt{4a-1}$, which yields a null derivative of the convergence rate of $\|\tilde{P}_t^{(a)} - \tilde{\mu}_a\|$ each times $kT_a, k \in \mathbb{N}$. One can also extend there results to the generator

$$\tilde{L}_{a,b,c,d} := by\partial_x - ax\partial_y + c\partial_y^2 - dy\partial_y$$

for which $\tilde{\mu}_{a,b,c,d} := \gamma_{bc/(ad)} \otimes \gamma_{c/d}$ is an invariant measure. The hypo-coercivity obtained is given by

$$\forall t \geqslant 0, \qquad \|\tilde{P}_t^{(a,b,c,d)} - \tilde{\mu}_{a,b,c,d}\|_{\mathbb{L}^2(\tilde{\mu}_{a,b,c,d})\circlearrowleft} = C_{ab/d^2}(dt) \exp\left(-\frac{1-\sqrt{(1-4abd^{-2})_+}}{2}dt\right).$$

As above, it is interesting to compare this rate with the one obtained by the semi-group $(\tilde{Q}_t^{(a,b,c,d)})_{t\geqslant 0}$ whose generator is $\tilde{K}_{a,b,c,d} := c\partial_x^2 - \frac{da}{b}x\partial_x$. This generator impulses the same amount of randomness per unit time as the hypo-coercive generator $\tilde{L}_{a,b,c,d}$ and $\tilde{K}_{a,b,c,d}$ is self adjoint for $\gamma_{bc/(ad)}$ (which is the marginal on coordinate x of $\tilde{\mu}_{a,b,c,d}$). After a rescaling step, $\tilde{K}_{a,b,c,d}$ is an Ornstein-Ulhenbeck generator with spectral gap da/b. The exponential convergence rate of $(\tilde{Q}_t^{(a,b,c,d)})_{t\geqslant 0}$ towards $\gamma_{bc/(ad)}$ is then da/b. Thus, if one chooses

$$\frac{a}{b} < \frac{1}{2}\left(1 - \sqrt{\left(1 - 4\frac{ab}{d^2}\right)_+}\right),$$

it is still better to use an hypo-coercive semi-group $(\tilde{P}_t^{(a,b,c,d)})_{t\geqslant 0}$ than the use of standard $(\tilde{Q}_t^{(a,b,c,d)})_{t\geqslant 0}$ for the simulation of $\gamma_{bc/(ad)}$. Hence, the same conclusion (as the one given in the paragraph above) still holds.

4.4 Average diffusion with small parameter

We now come back to the average gradient diffusion described by eqrefsde and focus on small perturbations of this dynamical system. We first tackle the problem of perturbations of trajectories and then study the question somewhat more intricate of perturbations of invariant measures. We shall restrict this problem to the homogeneous Markov case which corresponds to the memory function $k(t) = e^{\lambda t}$ described in the paragraph above :

$$\begin{cases} dX_t = \varepsilon dW_t - Y_t dt, \\ dY_t = \lambda(\nabla U(X_t) - Y_t)dt. \end{cases} \tag{4.20}$$

Remind that z will refer to the couple (x, y) as well as $(Z_t^\varepsilon)_{t\geqslant 0}$ will denote the coupled process $(X_t^\varepsilon, Y_t^\varepsilon)_{t\geqslant 0}$ with a level ε of noise associated to (4.20). In the sequel, we will denote by ν_ε the unique invariant measure of (4.20) (uniqueness is satisfied under assumption 11 denoted $\mathbf{H_{Hypo}}$), $(P_t^\varepsilon(z,.)$ will be its associated semi-group and at last $\mathcal{A}^\varepsilon$ is the generator of (4.20).

4.4.1 Large deviations of finite time trajectories

Of course, the limiting behaviour of (4.20) when $\varepsilon \to 0$ is strongly related to the behaviour of the deterministic dynamical system obtained when $\varepsilon = 0$ which is here

$$\begin{cases} \dot{x}(t) = -y(t). \\ \dot{y}(t) = \lambda(\nabla U(x(t)) - y(t)). \end{cases} \tag{4.21}$$

This link between (4.20) and (4.21) will be obtained through optimal solutions of the controlled problem already pointed for the minoration of the density $p_t(z_0, z)$. We define on $\mathbb{R}^d \times \mathbb{R}^d$ the drift vector field $b(z) = (-y, \lambda[\nabla U(x) - y])$, the controlled problem associated to (4.20) comes down to study for any $\varphi \in \mathbb{H}_0^1$ (which stands for the Cameron-Martin space) the behaviour of $\mathbf{z}_\varphi := (\mathbf{z}_\varphi(t))_{t\geqslant 0}$ and $\tilde{\mathbf{z}}_\varphi := (\tilde{\mathbf{z}}_\varphi(t))_{t\geqslant 0}$, that satisfy

$$\dot{\mathbf{z}}_\varphi := b(\mathbf{z}_\varphi) + \begin{pmatrix} \dot{\varphi} \\ 0 \end{pmatrix} \quad \text{et} \quad \dot{\tilde{\mathbf{z}}}_\varphi := -b(\tilde{\mathbf{z}}_\varphi) + \begin{pmatrix} \dot{\varphi} \\ 0 \end{pmatrix}. \tag{4.22}$$

Under the assumption 14 (denoted $(\tilde{\mathbf{H}}_U)$ in section 4.2), it is possible to show non explosion of controlled trajectories within a finite time horizon. Moreover, we establish a preliminary result of Large Deviation Principle (L.D.P.) within finite time, even if the diffusion is totally degenerated on coordinate y.

Proposition 4.4.1 *Assume that U satisfies assumption $(\tilde{\mathbf{H}}_U)$, then for any $z \in \mathbb{R}^d$ and all sequence $(z_\varepsilon)_{\varepsilon>0} \to z$ when $\varepsilon \to 0$, the Markov process $Z^\varepsilon = (X^{(\varepsilon)}, Y^{(\varepsilon)})$ satisfies a L.D.P. on $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^{2d})$ (endowed with the uniform convergence on compact sets topology). The rate is ε^{-2} and the good rate function $\mathcal{I}_z$ is defined for any absolutely continuous function $\mathbf{z} = (\mathbf{z}(t))_{t\geqslant 0}$ which satisfies $\mathbf{z}(0) = z$ by*

$$\mathcal{I}_z(\mathbf{z}) = \frac{1}{2} \inf_{\varphi \in \mathbb{H}_0^1 | \mathbf{z} = \mathbf{z}_\varphi} \int_0^\infty |\dot{\varphi}(s)|^2 ds.$$

In particular, for all $t \geqslant 0$ and $z \in \mathbb{R}^{2d}$, $(P_t^\varepsilon(z_\varepsilon, .))_{\varepsilon>0}$ satisfies a L.D.P. with rate ε^{-2} and good rate function $I_{z,t}$ defined for all $z, z' \in \mathbb{R}^{2d}$ by

$$I_{z,t}(z') = \inf_{\mathbf{z} \in \mathcal{Z}_t(z,z')} \mathcal{I}_z(\mathbf{z}) \tag{4.23}$$

where $\mathcal{Z}_t(z,z')$ *is the set of absolutely continuous trajectories* $\mathbf{z}$ *that leads* z *to* z' *in finite time* t.

The main difficulty for the proof of such proposition is to show a contraction principle, this is stronly related to the non explosion of controlled trajectories. This last point is obtained using a perturbation of Gronwall's Lemma for the Lyapunov function $\mathcal{E}(x,y) = U(x) + |y|^2/(2\lambda$. Remark that it may be viewed as an example of extensions of Schilder's Theorem which is generalized in [Azencott, 1980].

4.4.2 Large deviations sub-sequences of $(\nu_\varepsilon)_{\varepsilon\mapsto 0}$

The large deviation property on stationary measures $(\nu_\varepsilon)_{\varepsilon\mapsto 0}$ next depends on several conditions. The first one is an exponential tightness property on $(\nu_\varepsilon)_{\varepsilon\mapsto 0}$. This property is obtained by considering hitting times of compact sets for the process $(Z^\varepsilon_t)_{t\geqslant 0}$ when $\varepsilon \mapsto 0$. The expectation of functions of hitting times are also estimate using Lyapunov functions. In view of this tightness property, the standard Lyapunov function $\mathcal{E}(x,y)$ evoked above is not sufficient and one should use a function which globally controls both coordinates x and y in (4.20). The trick still comes to use an application built from $V(x,y)$ already defined (4.11), more precisely if we denote

$$\tilde{V}^\varepsilon(x,y) = \exp\left(\delta\varepsilon^{-2}V^p(x,y)\right),$$

a good tuning of coefficients δ and p permits to obtain the following contraction

$$\mathcal{A}^\varepsilon\tilde{V}^\varepsilon \leqslant \delta\varepsilon^{-2}(\beta - \alpha V^p).$$

We can then deduce the following result.

Proposition 4.4.2 *Assume that* U *is such that* $(\tilde{\mathbf{H}}_U)$ *holds, then there exists a compact* B *of* $\mathbb{R}^{2d}$*, such that the hitting time* τ_ε *of* B *satisfies :*

i) *For any compact* K, $\limsup_{\varepsilon\to 0}\sup_{z\in K}\mathbb{E}_z[(\tau_\varepsilon)^2] < \infty$.
ii) $\exists\,\delta$ *such that for any compact* K, $\limsup_{\varepsilon\to 0}\sup_{z\in K}\sup_t \mathbb{E}_z[|Z^{(\varepsilon)}_{t\wedge\tau_\varepsilon}|^{\frac{\delta}{\varepsilon^2}}]^{\varepsilon^2} < +\infty$.
iii) *For any compact* K *such that* $K\cap B = \varnothing$, $\liminf_{\varepsilon\to 0}\inf_{z\in K}\mathbb{E}_z[\tau_\varepsilon] > 0$.

We then can deduce the main result of this paragraph which establishes a L.D.P. up to a sub-sequence $(\varepsilon_n)_{n\in\mathbb{N}}$ and an Hamilton-Jacobi equation satisfied by the good rate function.

Theorem 4.4.1 *Assume that* $(\tilde{\mathbf{H}}_U)$ *holds, then* $(\nu_\varepsilon)_{\varepsilon\in(0,1]}$ *is exponentially tight. Moreover, for any sub-sequence* $(\varepsilon_n)_{n\in\mathbb{N}}$ *along which a L.D.P. holds*[6] *with rate* ε_n^{-2}*, the good rate function* W *satisfies*

$$\forall t\geqslant 0 \quad \forall z\in\mathbb{R}^d\times\mathbb{R}^d \qquad W(z) = \inf_{\begin{cases}\varphi\in\mathbb{H}^1_0\\ \mathbf{z}_\varphi(0)=z\end{cases}} \left[\frac{1}{2}\int_0^t |\dot{\varphi}|^2 + W(\tilde{\mathbf{z}}_\varphi(t))\right]. \tag{4.24}$$

This theorem provides only a partial existence of a L.D.P. for the sequence $(\nu_\varepsilon)_{\varepsilon\geqslant 0}$ since the obtained rate functions solutions of (4.24) may not be all the same. Indeed, this Hamilton-Jacobi equation (described here in the variational form of optimal control of dynamical programming principle) does not present some uniqueness property of its solution, and thus we cannot deduce any uniqueness property of W from (4.24). The main goal of the next paragraph is to provide sufficient conditions to obtain a L.D.P. along all the sequence $(\nu_\varepsilon)_{\varepsilon\geqslant 0}$.

6. Such sub-sequence will be refered as a LD-convergent sub-sequence

4.4.3 Freidlin & Wentzell estimates

Equilibrium of the vector field We now assume the main hypothesis which is necessary to obtain further results on the dynamical system (4.4).

Assumption 15 ($\mathbf{H_D}$) *The set of critical points of* U *is discrete (thus finite), and the Hessian of* U *is invertible on all these critical points.*

We will denote in the sequel $\{x_1^\star, \dots x_\ell^\star\}$ the set of these critical points of U. The elementary property which permits to identify equilibriums of (4.21) which uses the vector field $+b$ is as follows

Proposition 4.4.3 *Under the assumption* $(\mathbf{H_D})$*,equilibriums of* (4.21) *are* $\{z_1^\star, \dots, z_\ell^\star\} := \{(x_1^\star, 0), \dots, (x_{ell}^\star, 0)\}$*. The stable points are the ones for which* $x_i^\star$ *is a local minimum of* U*.*

Under this assumption $(\mathbf{H_D})$, it is possible to extend equation(4.24) to an infinite horizon, hence the good rate function W is indeed solution of :

$$\forall z \in \mathbb{R}^d \times \mathbb{R}^d \qquad W(z) = \min_{1 \leqslant i \leqslant \ell} \inf_{\begin{cases} \varphi \in \mathbb{H}_0^1 \\ \mathbf{z}_\varphi(0) = z, \quad z_\varphi(+\infty) = z_i^\star \end{cases}} \left[\frac{1}{2} \int_0^t |\dot{\varphi}|^2 + W(z_i^\star) \right]. \tag{4.25}$$

The proof of such equality relies on dynamical argument of the vector field $+b$: non explosion of trajectories in infinite horizon, compactness of trajectories and ω-limit sets. In particular, the proof of (4.25) still does not use any uniqueness argument for W. However, the important point in formula (4.25) is to remark that the function W defined on $\mathbb{R}^d \times \mathbb{R}^d$ depends exclusively on its values $W(z_i^\star)$ taken in equilibrium points of $+b$. These values $W(z_i^\star)$ are provided by the Freidlin & Wentzell estimates.

Freidlin & Wentzell theorey The idea is to exploit an explicit representation of the invariant measures $(\nu_\varepsilon)_{\varepsilon \geqslant 0}$ obtained through a skeleton chain built from the hitting and exiting times of neighbourhoods of equilibriums. We extend this due to [Has'minskii, 1980] (which is a key poinf of the approach of [Freidlin and Wentzell, 1984]), to our hypo-elliptic case using control arguments of trajectories to ensure that the skeleton representation corresponds to a *finite* measure proportional to ν_ε.

Proposition 4.4.4 *Let us denote* $\tilde{\mu}_\varepsilon$ *the unique invariant measure of the skeleton Markov chain that lives in* $\cup_{i=1}^\ell g_i$*, the the measure defined for any Borelian set* A *of* $\mathbb{R}^d \times \mathbb{R}^d$ *by*

$$\mu_\varepsilon(A) := \int_{\partial g} \tilde{\mu}_\varepsilon^{\partial g}(dz) \mathbb{E}_z \int_0^{\tau_1(\partial g)} \mathbf{1}_{Z_s^{z,\varepsilon} \in A} ds$$

is an invariant measure finite and proportional to the invariant distribution ν_ε*.*

Next, we show that Freidlin & Wentzell estimates can be applied to our skeleton chain. For any couple of points ξ_1 and ξ_2, we define the optimal control cost with finite time T to reach ξ_2 from ξ_1 as

$$I_T(\xi_1, \xi_2) := \inf_{\begin{cases} \varphi \in \mathbb{H}_0^1 \\ \mathbf{z}_\varphi(0) = \xi_1, \mathbf{z}_\varphi(T) = \xi_2 \end{cases}} \frac{1}{2} \int_0^T |\dot{\varphi}(s)|^2 ds,$$

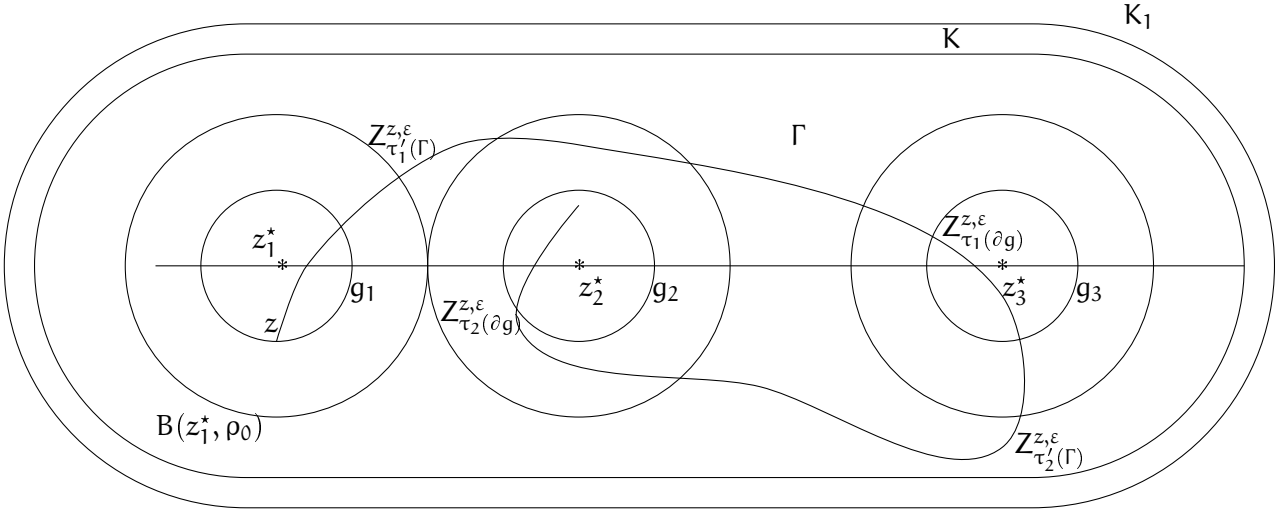


FIGURE 4.1 – Representation of neighbourhoods g_i of equilibriums points and process $(Z_t^{z,\varepsilon})_{t\geqslant 0}$, the skeleton chain is described by the transition of $Z^{z,\varepsilon}_{\tau_1(\partial g)} \mapsto Z^{z,\varepsilon}_{\tau_2(\partial g)}$ which belong to $\cup_{i=1}^{\ell} g_i$.

and the optimal control cost is $I(\xi_1,\xi_2) := \inf_{T\geqslant 0} I_T(\xi_1,\xi_2)$. In a similar way,

$$\tilde{I}(z_i^\star, z_j^\star) := \inf_{T>0} \inf\left\{\frac{1}{2}\int_0^T |\dot{\varphi}(s)|^2 ds, \varphi \in \mathbb{H}_0^1, \mathbf{z}_\varphi(0) = z_i^\star, \mathbf{z}_\varphi(T) = z_j^\star, \forall s \in [0,T],\ \mathbf{z}_\varphi(z_i^\star, s) \notin \cup_{k\neq i,j} g_k\right\}.$$

It is then possible to use a L.D.P. on trajectories of finite horizon and the exact local controllability near each $z_i^\star$ to obtain sharp approximations of the transitions of the skeleton chain using I when ε is small enough (see [20] and [Freidlin and Wentzell, 1984]) . We can prove the following result.

Proposition 4.4.5 *Assume that* U *satisfies the assumptions* $(\mathbf{H_D})$, $(\tilde{\mathbf{H}}_\mathbf{U})$ *and* $(\mathbf{H_{Hyp}})$, *then :*

i) *For any couple* $(i,j) \in \{1\dots\ell\}^2$, $\tilde{I}(z_i^\star, z_j^\star) < +\infty$ *and there exists only one communication class of* g_i *for the skeleton chain.*

ii) *Moreover, for any* $\gamma > 0$, *one can find* ρ_0 *and* ρ_1 *(size of neighbourhoods of* $z_i^\star$ *for the definition of the skeleton chain) such that* $0 < \rho_1 < \rho_0$ *and for which a sufficiently small* ε *yields*

$$\forall (i,j) \in \{1\dots\ell\}^2 \quad \forall x \in \partial g_i \qquad 0 < e^{-\varepsilon^{-2}[\tilde{I}(z_i^\star, z_j^\star)+\gamma]} \leqslant \tilde{\mathbb{P}}^\varepsilon(x, \partial g_j) \leqslant e^{-\varepsilon^{-2}[\tilde{I}(z_i^\star, z_j^\star)-\gamma]}.$$

4.4.4 Large Deviation Principle for invariant measures $(\nu_\varepsilon)_{\varepsilon\geqslant 0}$

The above estimation permits to compute a sharp approximation of the stationary measure of the skeleton Markov chain using the notion of $\{i\}$-Graphs. Remind briefly that for any $i \in \{1,\dots,\ell\}$, $\mathcal{G}(i)$ is the set of oriented graphs with vertices $\{z_1^\star,\dots,z_\ell^\star\}$ and such that

(i) All vertex $z_j^\star \neq z_i^\star$ is the starting point of exactly one edge.

(ii) The graph does not contain any cycle.

(iii) For any $z_j^\star$, there exists a unique path of oriented edges starting at $z_j^\star$ which reaches $z_i^\star$.

At last, the estimation of μ_ε for the skeleton Markov chain deduced from proposition 4.4.5 associated to the "link" formula of ν_ε given by proposition 4.4.4 permet alors de conclure le résultat suivant.

Theorem 4.4.2 *Under assumptions* $(\mathbf{H_{Hypo}}), (\mathbf{H_D})$ *and* $(\tilde{\mathbf{H}}_\mathbf{U})$, *for any* (ε_n) LD*-convergent sub-sequence, the good rate function* W *satisfies*

$$\forall i \in \{1 \dots \ell\} \qquad W(z_i^\star) = \min_{g \in \mathcal{G}(i)} \sum_{(z_m^\star \to z_n^\star) \in g} I(z_m^\star, z_n^\star) = \min_{g \in \mathcal{G}(i)} \sum_{(z_m^\star \to z_n^\star) \in g} \tilde{I}(z_m^\star, z_n^\star). \tag{4.26}$$

Moreover, W *is uniquely defined by* (4.26) *and*

$$\forall z \in \mathbb{R}^d \times \mathbb{R}^d \qquad W(z) = \min_{1 \leqslant i \leqslant \ell} \inf_{\begin{cases} \varphi \in \mathbb{H}_0^1 \\ \tilde{\mathbf{z}}_\varphi(0) = z, \tilde{\mathbf{z}}_\varphi(+\infty) = z_i^\star \end{cases}} \left[\frac{1}{2} \int_0^\infty |\dot{\varphi}|^2 + W(z_i^\star) \right],$$

thus $(\nu_\varepsilon)_{\varepsilon \geqslant 0}$ *satisfies a Large Deviation Principle.*

4.4.5 Quasi-potential for a double-well potential

The variation of the quasi-potential W (rate function of the L.D.P.) given in the paragraph above by Theorem 4.4.2 is simple when U is convex, but it is far from being also the case in a more general situation. We are going to study the particular case of a potential defined on $\mathbb{R}$ which is not convex and possesses a double well. This potential U is typically described in Figure 4.2.

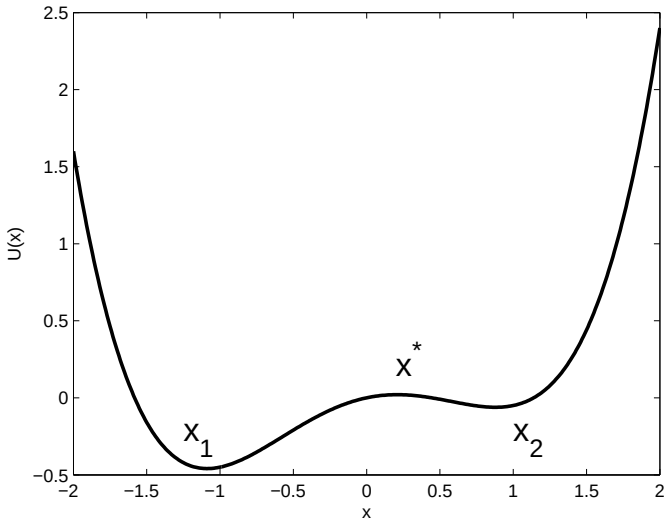


FIGURE 4.2 – Double-well potentiel U with 2 minima $x_1 < x_2$ and one local maximum x^*.

We then aim to compute the quasi-potential associated to the L.D.P. obtained for the process (4.4), its formal expression is quite simple since only L^2 costs of control to transit between $z_1^\star := (x_1, 0)$ and $z_2^\star := (x_2, 0)$ are necessary (owing to the simplicity of i-graphs). Without loss of generality, we assume that $U(x_1) < U(x_2)$ and we wish to compute a lower bound of $W(z_2^\star) = I(z_1^\star, z_2^\star)$ as well as an upper bound of $W(z_1^\star) = I(z_2^\star, z_1^\star)$.

Upper and lower bound in the standard case In standard case of simulated annealing, the drift term corresponds to the opposite of a gradient and the control problem is non degenerated and written as $\dot{z} = -\nabla U(z) + \varphi$. One may remark that the particular choice $\varphi(z) = 2\nabla U(z)$ enables to reach a local maxima $x^\star$ from a local minimum x_1 with a cost equals to $2[U(x^\star) - U(x_1)]$. Then, a continuity argument of the cost permits to obtain an identical cost between x_1 and x_2, and the following upper bound easily follows

$$I(x_1, x_2) \leqslant 2[U(x^\star) - U(x_1)].$$

Moreover, a simple argument lead to a matching lower bound : for any trajectory $(z_t)_{t\geqslant 0}$ starting at x_1 which leads to x_2 necessary reaches x^* [7]

$$|\dot{\varphi}|^2 = |\dot{z} + \nabla U|^2 \geqslant 2\langle \dot{z}, \nabla U(z)\rangle.$$

and we then obtain that the control cost is bounded from below by $2[U(x^\star) - U(x_1)]$.

Upper and lower bound in the average gradient system Indeed, it is possible to expand the former results to a slightly more general case of drifts (see for instance the works of [Sheu, 1986]) but the problem is largely open for a general drift given by $b(x, y) = (-y, \lambda(\nabla U(x) - y)$. In order to find a good trajectory which drives $z_2^\star$ to $z_1^\star$, we have been inspired by the standard case and we exploit the idea to "invert" the drift in view to go back in time. This is traduced by the following differential equation

$$dX_t = \frac{1}{e^{\lambda t}} \int_0^t \lambda e^{\lambda s} \nabla U(X_s) ds. \tag{4.27}$$

Since the control φ only acts on the first coordinate, it is quite natural to choose $\dot{\varphi} = 2y$ since it yields the desired differential equation described by (4.27). This method thus finds a trajectory with a cost identical to the one obtained in the standard case.

Proposition 4.4.6 *For the double-well potential described above, we have*

$$W(z_1^\star) = I(z_2^\star, z_1^\star) \leqslant 2[U(x^\star) - U(x_2)].$$

Finding a suitable lower bound of $W(z_2)$ is clearly a much more difficult task and may be tackled by considering either controlled trajectory with controls which act on x and y, or in a more natural way by limiting the control to act only on the x coordinate. We immediately remark that :

$$I_T(z_1^\star, z_2^\star) = \inf_{\begin{cases} \varphi \in \mathbb{H}_0^1 \\ \mathbf{z}_\varphi(0) = z_1^\star \\ \mathbf{z}_\varphi(T) = z_2^\star \end{cases}} \frac{1}{2}\int_0^T |\dot{\varphi}(s)|^2 ds \geqslant \inf_{\begin{cases} \varphi, \psi \in \mathbb{H}_0^1 \\ \mathbf{z}_{\varphi,\psi}(0) = z_1^\star \\ \mathbf{z}_{\varphi,\psi}(T) = z_2^\star \end{cases}} \frac{1}{2}\int_0^T |\dot{\varphi}(s)|^2 + |\dot{\psi}(s)|^2 ds$$

where $\mathbf{z}_{\varphi,\psi}$ is a x/y controlled trajectory by φ and ψ. In the sequel, we only provide the approach developed for the degenerated control problems which only acts on coordinate x, the other approach may be described in [20] and provides interesting results and quite more general results (on U) than the ones detailed here. Note that the lower bound here are better for the

7. In larger dimension, one should consider the minimal elevation necessary to climb the hill between x_1 and x_2

degenerate control but the needed assumptions are slightly restrictive. For any φ controlled trajectory , we have

$$|\dot{\varphi}|^2 = |\dot{x} + y|^2 = \dot{x}^2 + y^2 + 2\dot{x}y,$$

and we aim to bound this quadratic form in $(x, y, \dot{x})$ from below by the derivative along the trajectory z_φ of a suitable function of x and y. The principle is thus similar to the one used in the standard approach when we used $|\dot{\varphi}|^2 \geqslant 2\langle \dot{z}, \nabla U(z)\rangle$. Thus, we wish to find $\mathcal{L}(x, y)$ such that

$$\dot{x}^2 + y^2 + 2\dot{x}y \geqslant \langle \nabla \mathcal{L}(x, y), (\dot{x}, \dot{y})\rangle. \tag{4.28}$$

We researched map $\mathcal{L}$ is of the form

$$\mathcal{L}_{\alpha,\beta,\gamma}(x, y) := \alpha U(x) + \beta y^2/2 - \gamma y U'(x),$$

and it is quite intricate to remark that such function may be used both to obtain compactness results in large time, and also may yield some lower bound of $\mathbb{L}^2$ control cost between two points $z_1^\star$ and $z_2^\star$. We then obtain the following result.

Proposition 4.4.7 *For any $\alpha \in [0, 2]$, there exists an explicit $m(\alpha, \lambda)$ such that $\|U''\|_\infty \leqslant m(\alpha)$ implies that one can find $\beta(\alpha)$ and $\gamma(\alpha)$ so that (4.28) holds. For this choice, we have*

$$I_T(z_1^\star, z_2^\star) \geqslant \alpha[U(x^\star) - U(x_1)].$$

We will instantaneously remark that this proposition cannot reach a lower bound greater than twice the elevation of U between x_1 and x_2, which is coherent with the result of 4.4.6. These two results combined with Theorem 4.4.2 permits to give the final result on the behaviour of ν_ε towards the global minimum of U.

Theorem 4.4.3 *Under assumptions $(\mathbf{H_{Hypo}})$, $(\mathbf{H_D})$ and $(\tilde{\mathbf{H}}_\mathbf{U})$, if U is a double-well real valued potential (described as above) with $U(x_1) < U(x_2)$ and such that $\|U''\|_\infty \leqslant m\left(2\frac{U(x^\star)-U(x_2)}{U(x^\star)-U(x_1)}, \lambda\right)$, then*

$$\lim_{\varepsilon \mapsto 0} \nu_\varepsilon = \delta_{x_1}.$$

At last, remark that when λ comes larger, the average gradient system uses a shorter range memory and the bound on $\|U''\|_\infty$ in the former result is much more permissive. This phenomenon is illustrated in Figure 4.3 which shows for several values of λ the evolution of α which is the multiplicative coefficient of $U(x^\star) - U(x_1)$ that depends on $\|U''\|_\infty$.

4.5 Further developments

4.5.1 Hypo-coercivity of the memoru gradient diffusion, simulated annealing

The first forthcoming work would concern some hypo-coercivity results and thus how obtain an upper bound of its semi-group in $\mathbb{L}^2(\nu)$. This points has not been addressed since our results was obtained using [Down et al., 1995] techniques, which yield total variation results. It would be interesting to find a stronger result. An incidence angle to go towards such result should certainly use the Lyapunov function in order to obtain some Poincaré-like inequality. Such method has already been exploited in the case of kinetic Fokker-Planck equations (see for instance a brief exposition of such method in [Villani, 2006]). An additional difficulty in the case of memory gradient diffusion is the non explicit nature of its invariant distribution and only implicit relations are known through a P.D.E. satisfied by ν.

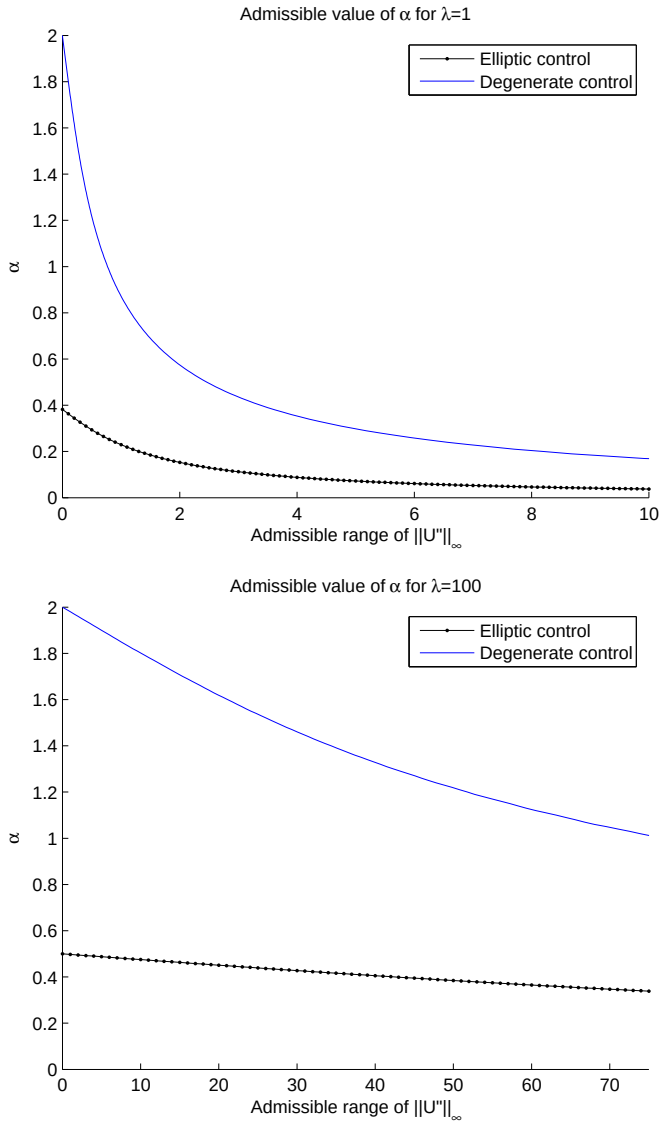


FIGURE 4.3 – Evolution of α multiplicative coefficient of the elevation of U with respect to $\|U''\|_\infty$ for several λ.

The second natural point would continue the annealing study in view of a real simulated annealing procedure, ε should now become an evanescent function of t. Numerical experiments (not shown in this memory) has pointed that it was possible to use a temperature scheme $\varepsilon(t) = c/\log t$ and obtain a convergence of $(Z_t^{\varepsilon(t)})_{t\geqslant 0}$ towards the global minimum of U in the case of a double-well potential, provided that the "hill" between the two wells is enough undershot (see assumptions of Theorem 4.4.3). Moreover, there still also exists other numerical hints which would prove that the constant c may be chosen lower than the limiting one in the case of the standard simaulted annealing procedure. At last, an optimisation with respect to λ for the memory simulated annealing seems to be important. There is no theoretical response on all such points at the moment.

4.5.2 Controllability result on the memory system

Another important class of problem is the nature of controllability results we may obtain for the system 4.22. We show in our study that under hypothesis of non degeneracy near critical points of U and growing conditions at ∞, the approached controllability is true. Even if the growing condition seems imperative on U, it seems to be more discussable regarding the assumption of non degeneracy. Indeed, there exists a large amount of methods to avoid the use of the Kalman linearisation method, see for instance Sussman conditions in [Coron, 2007] to obtain local exact controllability, or fixed point techniques (an example may be found in [Beauchard and Zuazua, 2009]).

At last, from a numerical point of view, it seems challenging to develop algorithms for the computation of optimal control costs to obtain W. A starting collaboration with numerical specialists of controllability has lead us to consider the Pontryagin maximum principle to obtain numerical results.

4.5.3 Non reversible simulations

Conclusions drawn by the paragraph 4.3 should stimulate the interest of second order simulations (second order Markov chains, kinetic equations) in order to obtain faster convergence rates to steady regime than the one obtained by first order dynamical systems. This point is not true of course in full generality, and should be completed by generic examples. This is not the case at the moment regarding for instance the very partial results obtained in the case of kinetic Fokker-Planck equations. If such phenomenon holds, it would be of first interest for stochastic algorithms which uses MCMC simulations or Bayesian algorithms.

Bibliography

Thesis

[1] Sébastien Gadat. Apprentissage d'un vocabulaire symbolique pour la détection d'objets dans une image. *Thèse de l'École Normale Supérieure de Cachan*, 2004.

Published or accepted papers - Statistics in large dimensions

[2] Kim-Anh Lê Cao, Philippe Besse, Olivier Gonçalves, and Sébastien Gadat. Selection of biologically relevant genes with a wrapper stochastic algorithm. *Statistical Applications in Genetics and Molecular Biology*, 6, 2007.

[3] Kim-Anh Lê Cao, Agnès Bonnet, and Sébastien Gadat. Multiclass classification and gene selection with a stochastic algorithm. *Computational Statistics and Data Analysis*, 53 :3601–3615, 2009.

[4] Serge Cohen, Sébastien Déjean, and Sébastien Gadat. Adaptive sequential design for regression on multi-resolution bases. *Statistics and Computing, to appear*, 22(2) :1–20, 2012.

[5] Sébastien Gadat. Jump diffusion over feature space for object recognition. *Siam, Journal on Control and Optimisation*, 47 :904–935, 2008.

[6] Sébastien Gadat and Laurent Younes. A stochastic algorithm of features extraction for pattern recognition. *Journal of Machine Learning Research*, 8 :509–547, 2007.

[7] N. Villa, T. Dkaki, S. Gadat, J.M. Inglebert, and Q.D. Truong. Recherche et représentation de communautés dans un grand graphe : une approche combinée. *Document Numérique*, 14 :59–80, 2011.

Published or accepted papers - Deformable models

[8] J. Bigot, C. Christophe, and S. Gadat. Random action of compact lie groups and minimax estimation of a mean pattern. *IEEE, Transactions on Information Theory, to appear*, 2012.

[9] Jérémie Bigot and Sébastien Gadat. A deconvolution approach to estimation of a common shape in a shifted curves model. *Annals of Statistics*, 38(4) :2422–2464, 2010.

[10] Jérémie Bigot and Sébastien Gadat. Smoothing under diffeomorphic constraints with homeomorphic splines. *Siam, Journal on Numerical Analysis*, 48(1) :224–243, 2010.

[11] Jérémie Bigot, Sébastien Gadat, and Jean-Michel Loubes. Statistical m-estimation and consistency in large deformable models for image warping. *Journal of Mathematical Imaging and Vision*, 34(3) :270–290, 2009.

[12] Jérémie Bigot, Sébastien Gadat, and Clément Marteau. Sharp template estimation in a shifted curves model. *Electronic Journal of Statistics*, 4 :994–1021, 2010.

Published or accepted papers - Random dynamical systems

[13] A. Cabot, H. Engler, and S. Gadat. On the long time behavior of second order differential equations with asymptotically small dissipation. *Transactions of the American Mathematical Society*, 361 :5983–6017, 2009.

[14] A. Cabot, H. Engler, and S. Gadat. Second order differential equations with asymptotically small dissipation and piecewise flat potentials. *Electronic Journal of Differential Equations*, 17 :33–38, 2009.

[15] S. Gadat and F. Panloup. Long time behavior and stationary regime of memory gradient diffusions. *in revision for Annales de l'Institut Henri Poincaré (B)*, pages 1–40, 2012.

Submitted papers

[16] J. Bigot, S. Gadat, T. Klein, and C. Marteau. Intensity estimation of non-homogeneous poisson processes from shifted trajectories. *Preprint*, 2011.

[17] C. Cierco, M. Champion, S. Gadat, and M. Vignes. A boost-boost algorithm for high dimensional multivariate regression. *Preprint*, 2012.

[18] C. Cierco, M. Champion, S. Gadat, and M. Vignes. Gene network recovery and $\mathbb{L}^2$ boosting algorithm. *Preprint*, 2012.

[19] S. Gadat and L. Miclo. Spectral decompositions and l^2-operator norms of toy hypocoercive models. *Preprint*, 2012.

[20] S. Gadat, F. Panloup, and C. Pellegrini. Large deviation principle for invariant distributions of memory gradient diffusions. *Preprint*, 2012.

Books chapter

[21] J. Bigot and S. Gadat. *chapter : Pattern recognition through large deformations of images, in book Pattern Recognition.* Intech, 2010.

[22] S. Gadat. *chapter : Feature Selection in high dimension for face Detection, in book Advances in Face Image Analysis.* Techniques and Technologies, IGI - Global, 2009.

[23] J. Vandel D. Allouche C. Cierco-Ayrolles T. Schiex B. Mangin S. Gadat S. de Givry M. Vignes, M. Champion. *chapter : Integration of complementary approaches to reconstruct gene regulatory networks in a genetical genomics framework, in book Verification of methods for gene network inference from Systems Genetics data.* Springer, 2012.

Proceedings

[24] J.M. Azais, D. Debailleux, S. Gadat, and N. Suard. Assessment of an ionosphere storm occurrence risk. In *Proceedings of the 2011 Conference ENC GNSS*, London, England, November 2011.

[25] J.M. Azais, S. Gadat, C. Mercadier, and N. Suard. Gnss integrity achievement by using extreme value theory. In *Proceedings of the 2009 Conference ION GNSS*, San diego, USA, July 2009.

[26] J.M. Azais, S. Gadat, and N. Suard. Ionosphere severe storms and occurrence risk estimation. In *Proceedings of the 7th Conference Extreme Value Analysis, Probabilistic and Statistical Models and their Applications*, Lyon, France, June 2011.

[27] K.A. Lê Cao, S. Gadat, P. Besse, and O. Gonçalves. Application of a stochastic algorithm for gene selection. In *5th Workshop of Statistical methods for post-genomic data, 2007*, Paris, France, 2007.

[28] S. Gadat. Extraction of attributes for visual object recognition and dna microarray analysis. In *IEEE Workshop on Statistical Signal Processing, Bordeaux,. 2005*, Bordeaux, France, July 2005.

[29] S. Gadat. Reflected jump-diffusion for genes selection and classification of micro-array data. In *Workshop on Statistical Analysis of Postgenomic Data, 2005*, Paris, France, April 2005.

[30] S. Gadat. Sélection de variables pour la reconnaissance de formes. In *GRETSI'05 On Image and Signal treatment, 2005*, Louvain-La-Neuve, Belgique, September 2005.

[31] S. Gadat. Markov hybrid process for variable selection in classification. In *Proceedings of the 47th Conference on Decision and Control*, Cancun, Mexico, December 2008.

[32] S. Gadat. Bayesian consitency for deformable models in image processing. In *Proceedings of the 3th Annual Conference of Mathématiques pour l'Image.*, Orléans, France, June 2012.

[33] S. Gadat, O. Gonçalves, and K.A. Lê Cao. Gene selection with a stochastic algorithm for multiclass classification. In *In Proceedings of the 20th Annual Conference Proceedings of the 47th Conference Statistics for Data Mining, Learning and Knowledge Extraction.*, Aveiro, Portugal, August 2007.

[34] N. Villa, T. Dkaki, S. Gadat, J.M. Inglebert, and Q.D. Truong. Recherche et représentation de communautés dans des grands graphes. In *Proceedings of VSST 2009*, Nancy, France, 2009.

Technical reports

[35] J.M. Azais and S. Gadat. Automatisation de l'estimation par valeurs extremes pour la mesure d'intégrité. Technical report, Institut de Mathématiques de Toulouse, 2011.

[36] J.M. Azais, S. Gadat, A. Lagnoux, and C. Mercadier. Algorithmes de splitting pour la mesure d'intégrité. Technical report, Institut de Mathématiques de Toulouse, Institut Camille Jordan Lyon I, 2010.

[37] J.M. Azais, S. Gadat, and C. Mercadier. étude de la mesure d'integrite par la methode des valeurs extremes. Technical report, Institut de Mathématiques de Toulouse, Institut Camille Jordan Lyon I, 2009.

References

[A. Benveniste and Priouret, 1987] A. Benveniste, M. M. and Priouret, P. (1987). *Adaptive algorithms and stochastic approximations.*, volume 22 of *Applications of Mathematics*. Springer-Verlag, Berlin, Heidelberg, New York.

[Ait-Sahalia and Duarte, 2003] Ait-Sahalia, Y. and Duarte, J. (2003). Nonparametric option pricing under shape restrictions. *Journal of Econometrics*, 116 :9–47.

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19 :716–723. System identification and time-series analysis.

[Allassonière et al., 2007] Allassonière, S., Amit, Y., and Trouvé, A. (2007). Toward a coherent statistical framework for dense deformable template estimation. *Journal of the Statistical Royal Society (B)*, 69 :3–29.

[Allassonière et al., 2009] Allassonière, S., Kuhn, E., and Trouvé, A. (2009). Bayesian deformable models building via stochastic approximation algorithm : a convergence study. *Bernoulli*, 16 :641–678.

[Allon et al., 2007] Allon, G., Beenstock, M., Hackman, S., Passy, U., and Shapiro, A. (2007). Nonparametric estimation of concave production technologies by entropy. *Journal of Applied Econometrics*, 22 :795–816.

[Amit and Geman, 1997] Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7) :1545 ?1588.

[Antipin, 1994] Antipin, A. (1994). Minimization of convex functions on convex sets by means of differential equations (in russian). *Differential Equations*, 30 :1365–1375.

[Azencott, 1980] Azencott, R. (1980). *Large Deviations theory and Applications*, volume 774 of *Saint-Flour Summer school on Probability Theory*. Springer-Verlag.

[Bakry et al., 2008] Bakry, D., Cattiaux, P., and Guillin, A. (2008). Rate of convergence for ergodic continuous Markov processes : Lyapunov versus Poincaré. *J. Funct. Anal.*, 254(3) :727–759.

[Bakry and Émery, 1985] Bakry, D. and Émery, M. (1985). Diffusions hypercontractives. In *Séminaire de probabilités, XIX, 1983/84*, volume 1123 of *Lecture Notes in Math.*, pages 177–206. Springer, Berlin.

[Bally and Kohatsu-Higa, 2010] Bally, V. and Kohatsu-Higa, A. (2010). Lower bounds for densities of Asian type stochastic differential equations. *J. Funct. Anal.*, 258(9) :3134–3164.

[Barron et al., 1999] Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3) :301–413.

[Beauchard and Zuazua, 2009] Beauchard, K. and Zuazua, E. (2009). Some controllability results for the kolmogorov equation. *Ann. I. H. Poincaré, Analyse non linéaire*, 26 :1793–1815.

[Beirlant et al., 1999] Beirlant, J., Dierckx, G., Goegebeur, Y., and Matthys, G. (1999). Tail index estimation and an exponential regression model. *Extremes*, 2(2) :177–200.

[Ben Arous and Léandre, 1991] Ben Arous, G. and Léandre, R. (1991). Décroissance exponentielle du noyau de la chaleur sur la diagonale. II. *Probab. Theory Related Fields*, 90(3) :377–402.

[Ben Hassen and Haraux, 2011] Ben Hassen, I. and Haraux, A. (2011). Convergence and decay estimates for a class of second order dissipative equations involving a non-negative potential energy. *J. Funct. Anal.*, 260(10) :2933–2963.

[Benaim, 1996] Benaim, M. (1996). A dynamical system approach to stochastic approximations. *SIAM J. Control Optim.*, 34(2) :437–472.

[Benaïm and Hirsh, 1996] Benaïm, M. and Hirsh, M. (1996). Asymptotic pseudotrajectories and chain recurrent flows, with applications. *J. Dynam. Differential Equations*, 8 :141–176.

[Benaïm et al., 2002] Benaïm, M., Ledoux, M., and Raimond, O. (2002). Self-interacting diffusions. *Probab. Theory Related Fields*, 122 :1–41.

[Benveniste et al., 1990] Benveniste, A., Métivier, M., and Priouret, P. (1990). *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin. Translated from the French by Stephen S. Wilson.

[Bhattacharya and Patrangenaru, 2003] Bhattacharya, R. and Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds (i). *Annals of statistics*, 31(1) :1–29.

[Bhattacharya and Patrangenaru, 2005] Bhattacharya, R. and Patrangenaru, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds (ii). *Annals of statistics*, 33 :1225–1259.

[Bi et al., 2003] Bi, J., Bennett, K., Embrechts, M., Breneman, C., and Song, M. (2003). Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3 :1229–1243.

[Biau et al., 2008] Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, 9 :2015–2033.

[Bickel et al., 1998] Bickel, P. J., Ritov, Y., and Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Statist.*, 26(4) :1614–1635.

[Bickel et al., 2009] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4) :1705–1732.

[Bigot et al., 2009] Bigot, J., Gamboa, F., and Vimond, M. (2009). Estimation of translation, rotation, and scaling between noisy images using the Fourier-Mellin transform. *SIAM J. Imaging Sci.*, 2(2) :614–645.

[Bigot et al., 2010] Bigot, J., Loubes, J., and Vimond, M. (2010). Semiparametric estimation of shifts on compact lie groups for image registration. *Probability Theory and Related Fields*, pages 1–49.

[Binev et al., 2005] Binev, P., Cohen, A., Dahmen, W., DeVore, R., and Temlyakov, V. (2005). Universal algorithms for learning theory. I. Piecewise constant functions. *J. Mach. Learn. Res.*, 6 :1297–1321.

[Birgé, 1986] Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Relat. Fields*, 71(2) :271–291.

[Bissantz et al., 2007] Bissantz, N., Hohage, T., Munk, A., and Ruymgaart, F. (2007). Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.*, 45(6) :2610–2636 (electronic).

[Biswas and Chaudhuri, 2002] Biswas, A. and Chaudhuri, P. (2002). An efficient design for model discrimination and parameter estimation in linear models. *Biometrika*, 89(3) :709–718.

[Breiman, 1995] Breiman, L. (1995). Better subset selection using the non-negative garotte. *Technometrics*, 37(37) :738–754. With discussion, and a rejoinder by the authors.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1) :5–32.

[Breiman et al., 1984] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA.

[Bretagnolle and Huber, 1979] Bretagnolle, J. and Huber, C. (1979). Estimation des densités : risque minimax. *Z. Wahrsch. Verw. Gebiete*, 47(2) :119–137.

[Bühlmann, 2006] Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.*, 34(2) :559–583.

[Bühlmann and Yu, 2003] Bühlmann, P. and Yu, B. (2003). Boosting with the L_2 loss : regression and classification. *J. Amer. Statist. Assoc.*, 98(462) :324–339.

[Cabot, 2009] Cabot, A. (2009). Asymptotics for a gradient system with memory term. *Proc. Amer. Math. Soc.*, 137(9) :3013–3024.

[Candes and Tao, 2007] Candes, E. and Tao, T. (2007). The dantzig selector : statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6) :2313–2351.

[Cappé et al., 2005] Cappé, O., Moulines, E., and Ryden, T. (2005). *Inference in Hidden Markov Models*. Springer series in statistics. Springer Verlag, Paris.

[Carroll and Hall, 1988] Carroll, R. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, 83 :1184–1186.

[Castillo and van der Vaart A., 2012] Castillo, I. and van der Vaart A. (2012). Needles and straw in a haystack : posterior concentration for possibly sparse sequences. *preprint*.

[Cattiaux, 1992] Cattiaux, P. (1992). Stochastic calculus and degenerate boundary value problems. *Ann. Inst. Fourier*, 42 :541–624.

[Cavalier et al., 2002] Cavalier, L., Golubev, G. K., Picard, D., and Tsybakov, A. B. (2002). Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3) :843–874. Dedicated to the memory of Lucien Le Cam.

[Cavalier and Hengartner, 2005] Cavalier, L. and Hengartner, N. W. (2005). Adaptive estimation for inverse problems with noisy operators. *Inverse Problems*, 21(4) :1345–1361.

[Cavalier and Koo, 2002] Cavalier, L. and Koo, J.-Y. (2002). Poisson intensity estimation for tomographic data using a wavelet shrinkage approach. *IEEE Trans. Inform. Theory*, 48(10) :2794–2802.

[Cavalier and Raimondo, 2007] Cavalier, L. and Raimondo, M. (2007). Wavelet deconvolution with noisy eigenvalues. *IEEE Trans. Signal Process.*, 55(6, part 1) :2414–2424.

[Chiang et al., 1987] Chiang, T.-S., Hwang, C.-R., and Sheu, S. J. (1987). Diffusion for global optimization in $\mathbf{R}^n$. *SIAM J. Control Optim.*, 25(3) :737–753.

[Coppersmith and Diaconis, 1987] Coppersmith, D. and Diaconis, P. (1987). Random walk with reinforcement. *Unpublished*.

[Coron, 2007] Coron, J.-M. (2007). *Control and nonlinearity*, volume 136 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI.

[Cranston and Le Jan, 1995] Cranston, M. and Le Jan, Y. (1995). Self-attracting diffusions : two case studies. *Math. Ann.*, 303(1) :87–93.

[Davis et al., 1994] Davis, G., Mallat, S., and Zhang, Z. (1994). Adaptive time-frequency approximations with matching pursuits. In *Wavelets : theory, algorithms, and applications (Taormina, 1993)*, volume 5 of *Wavelet Anal. Appl.*, pages 271–293. Academic Press, San Diego, CA.

[de Haan and Ferreira, 2006] de Haan, L. and Ferreira, A. (2006). *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York. An introduction.

[de Sousa and Michailidis, 2004] de Sousa, B. and Michailidis, G. (2004). A diagnostic plot for estimating the tail index of a distribution. *J. Comput. Graph. Statist.*, 13(4) :974–995.

[Delarue and Menozzi, 2010] Delarue, F. and Menozzi, S. (2010). Density estimates for a random noise propagating through a chain of differential equations. *J. Funct. Anal.*, 259(6) :1577–1630.

[Desvillettes and Villani, 2001] Desvillettes, L. and Villani, C. (2001). On the trend to global equilibrium in spatially inhomogeneous entropy-dissipating systems : the linear Fokker-Planck equation. *Comm. Pure Appl. Math.*, 54(1) :1–42.

[Desvillettes and Villani, 2005] Desvillettes, L. and Villani, C. (2005). On the trend to global equilibrium for spatially inhomogeneous kinetic systems : the Boltzmann equation. *Invent. Math.*, 159(2) :245–316.

[Dette et al., 2006] Dette, H., Neumeyer, N., and Pilz, K. F. (2006). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli*, 12(3) :469–490.

[Dette and Pilz, 2006] Dette, H. and Pilz, K. F. (2006). A comparative study of monotone nonparametric kernel estimates. *J. Stat. Comput. Simul.*, 76(1) :41–56.

[Dette and Studden, 1997] Dette, H. and Studden, W. J. (1997). *The theory of canonical moments with applications in statistics, probability, and analysis*. Wiley Series in Probability and Statistics : Applied Probability and Statistics. John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

[DeVore and Temlyakov, 1996] DeVore, R. A. and Temlyakov, V. N. (1996). Some remarks on greedy algorithms. *Adv. Comput. Math.*, 5(2-3) :173–187.

[Diaconis et al., 2010a] Diaconis, P., Miclo, L., and Zuniga, J. (2010a). On the spectral analysis of second-order Markov chains. *Preprint*.

[Diaconis et al., 2010b] Diaconis, P., Miclo, L., and Zuñiga., J. (2010b). On the spectral analysis of second-order Markov chains. *Unpublished preprint*.

[Dolbeault et al., 2009] Dolbeault, J., Mouhot, C., and Schmeiser, C. (2009). Hypocoercivity for kinetic equations with linear relaxation terms. *C. R. Math. Acad. Sci. Paris*, 347(9-10) :511–516.

[Donoho, 1995] Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Trans. Inform. Theory*, 41(3) :613–627.

[Donoho et al., 2006] Donoho, D. L., Elad, M., and Temlyakov, V. N. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1) :6–18.

[Donoho et al., 2007] Donoho, D. L., Elad, M., and Temlyakov, V. N. (2007). On Lebesgue-type inequalities for greedy approximation. *J. Approx. Theory*, 147(2) :185–195.

[Donoho and Johnstone, 1994] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3) :425–455.

[Donoho and Johnstone, 1995] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432) :1200–1224.

[Donoho and Johnstone, 1998] Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3) :879–921.

[Donoho et al., 1995] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage : asymptopia ? *J. Roy. Statist. Soc. Ser. B*, 57(2) :301–369. With discussion and a reply by the authors.

[Douc et al., 2009] Douc, R., Fort, G., and Guillin, A. (2009). Subgeometric rates of convergence of f-ergodic strong markov processes. *Stochastic Processes and their Applications*, 119 :897–923.

[Down et al., 1995] Down, D., Meyn, S., and Tweedie, R. (1995). Exponential and uniform ergodicity of markov processes. *The Annals of Probability*, 23 :1671–1691.

[Drees and Kaufmann, 1998] Drees, H. and Kaufmann, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Process. Appl.*, 75(2) :149–172.

[Duflo, 1997] Duflo, M. (1997). *Random iterative models*, volume 34 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin. Translated from the 1990 French original by Stephen S. Wilson and revised by the author.

[Dupuis and Ramanan, 1999] Dupuis, P. and Ramanan, K. (1999). Convex duality and the skorokhod problem i, ii. *Probability Theory and Related Fields*, 115(2) :153–236.

[Durrett and Rogers, 1992] Durrett, R. and Rogers, L. (1992). Asymptotic behavior of brownian polymers. *Probab. Theory Related Fields*, 3 :337–349.

[Eckmann and Hairer, 2003] Eckmann, J.-P. and Hairer, M. (2003). Spectral properties of hypoelliptic operators. *Comm. Math. Phys.*, 235(2) :233–253.

[Efron et al., 2004] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32(2) :407–499. With discussion, and a rejoinder by the authors.

[Fedorov, 1972] Fedorov, V. V. (1972). *Theory of optimal experiments*. Academic Press, New York. Translated from the Russian and edited by W. J. Studden and E. M. Klimko, Probability and Mathematical Statistics, No. 12.

[Frechet, 1948] Frechet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de L'Institut Henri Poincaré*, 10 :215–310.

[Freidlin and Wentzell, 1984] Freidlin, M. I. and Wentzell, A. D. (1984). *Random perturbations of dynamical systems*, volume 260 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, New York. Translated from the Russian by Joseph Szücs.

[Freund and Schapire, 1997] Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.*, 55(1, part 2) :119–139. Second Annual European Conference on Computational Learning Theory (EuroCOLT '95) (Barcelona, 1995).

[Fromont et al., 2011] Fromont, M., Laurent, B., and Reynaud-Bouret (2011). Adaptive test of homogeneity for a poisson process. *Ann. Inst. H. Poincaré Probab. Statist.*, 47(1) :176–213.

[Fréchet, 1948] Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H.Poincaré, Sect. B, Prob. et Stat.*, 10 :235–310.

[Gamboa et al., 2007a] Gamboa, F., Loubes, J.-M., and Maza, E. (2007a). Semi-parametric estimation of shifts. *Electron. J. Stat.*, 1 :616–640.

[Gamboa et al., 2007b] Gamboa, F., Loubes, J.-M., and Maza, E. (2007b). Semi-parametric estimation of shifts. *Electron. J. Stat.*, 1 :616–640.

[Gasser and Kneip, 1992] Gasser, T. and Kneip, A. (1992). Statistical tools to analyze data representing a sample of curves. *Annals of Statistics*, 20(3) :1266–1305.

[Gasser and Kneip, 1995] Gasser, T. and Kneip, A. (1995). Searching for structure in curve samples. *Journal of the American Statistical Association*, 90(432) :1179–1188.

[Ghosal, 2000] Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.*, 74(1) :49–68.

[Ghosal et al., 2008] Ghosal, S., Lember, J., and van der Vaart, A. (2008). Nonparametric Bayesian model selection and averaging. *Electron. J. Stat.*, 2 :63–89.

[Ghosal and van der Vaart, 2001] Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5) :1233–1263.

[Grenander, 1993a] Grenander, U. (1993a). *General pattern theory - A mathematical study of regular structures.* Clarendon Press, Oxford.

[Grenander, 1993b] Grenander, U. (1993b). *General pattern theory - A mathematical study of regular structures.* Clarendon Press, Oxford.

[Grenander and Miller, 2007] Grenander, U. and Miller, M. (2007). *Pattern Theory : From Representation to Inference.* Oxford Univ. Press, Oxford.

[Guyon et al., 2006] Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. (2006). *Feature Extraction, Foundations and Applications.* Series Studies in Fuzziness and Soft Computing, Springer Verlag.

[Guyon et al., 2002] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46 :389–422.

[Hairer, 2011] Hairer, M. (2011). On malliavin's proof of hörmander's theorem. *Bull. Sci. Math.*, 165.

[Hale, 1988] Hale, J. K. (1988). *Asymptotic behavior of dissipative systems*, volume 25 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI.

[Hall and Huang, 2001] Hall, P. and Huang, L.-S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics*, 29 :624–647.

[Haraux, 1991] Haraux, A. (1991). *Systèmes dynamiques dissipatifs et applications*, volume 17 of *Recherches en Mathématiques Appliquées [Research in Applied Mathematics].* Masson, Paris.

[Haraux, 2007] Haraux, A. (2007). Sharp estimates of bounded solutions to some second-order forced dissipative equations. *J. Dynam. Differential Equations*, 19(4) :915–933.

[Has'minskii, 1980] Has'minskii, R. (1980). *Stochastic stability of differential equations.* Sijthoff & Noordhoff, Alphen aan den Rijn (The Nederlands).

[Has′minskiĭ and Ibragimov, 1990] Has′minskiĭ, R. and Ibragimov, I. (1990). On density estimation in the view of Kolmogorov's ideas in approximation theory. *Ann. Statist.*, 18(3) :999–1010.

[Helffer and Nier, 2005] Helffer, B. and Nier, F. (2005). *Hypoelliptic estimates and spectral theory for Fokker-Planck operators and Witten Laplacians*, volume 1862 of *Lecture Notes in Mathematics.* Springer-Verlag, Berlin.

[Hérau and Nier, 2004] Hérau, F. and Nier, F. (2004). Isotropic hypoellipticity and trend to equilibrium for the Fokker-Planck equation with a high-degree potential. *Arch. Ration. Mech. Anal.*, 171(2) :151–218.

[Hoerl and Kennard, 1975] Hoerl, A. E. and Kennard, R. W. (1975). A note on a power generalization of ridge regression. *Technometrics*, 17 :269.

[Hoffmann et al., 2006] Hoffmann, A., Siem, A., den Hertog, D., Kaanders, J., and Huizenga, H. (2006). Derivative-free generation and interpolation of convex pareto optimal imrt plans. *Physics in Medicine and Biology*, 51 :6349–6369.

[Hörmander, 1967] Hörmander, L. (1967). Hypoelliptic second order differential equations. *Acta Mathematica*, 117 :147–171.

[Ibragimov and Has′minskiĭ, 1981] Ibragimov, I. A. and Has′minskiĭ, R. Z. (1981). *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York. Asymptotic theory, Translated from the Russian by Samuel Kotz.

[Johnstone et al., 2004] Johnstone, I., Kerkyacharian, G., Picard, D., and Raimondo, M. (2004). Wavelet deconvolution in a periodic setting. *J. Roy. Statist. Soc. Ser. B*, 66 :547–573.

[Joshi et al., 2004] Joshi, S., Davis, B., Jomier, B. M., and B, G. G. (2004). Unbiased diffeomorphic atlas construction for computational anatomy. *Neuroimage*, 23 :151–160.

[Kendall, 1984] Kendall, D. (1984). Shape manifolds, procrustean metrics, and complex projective spaces. *Bull. London Math Soc.*, 16 :81–121.

[Kiefer and Wolfowitz, 1959] Kiefer, J. and Wolfowitz, J. (1959). Optimum designs in regression problems. *Ann. Math. Statist.*, 30 :271–294.

[Kim, 1998] Kim, P. T. (1998). Deconvolution density estimation on SO(N). *Ann. Statist.*, 26(3) :1083–1102.

[Kneip and Gasser, 1988] Kneip, A. and Gasser, T. (1988). Convergence and consistency results for self-modelling regression. *Annals of Statistics*, 16 :82–112.

[Kohn, 1978] Kohn, J. J. (1978). Lectures on degenerate elliptic problems. In *Pseudodifferential operator with applications (Bressanone, 1977)*, pages 89–151. Liguori, Naples.

[Kolaczyk, 1999] Kolaczyk, E. D. (1999). Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds. *Statist. Sinica*, 9(1) :119–135.

[Koo and Kim, 2008] Koo, J. Y. and Kim, P. T. (2008). Asymptotic minimax bounds for stochastic deconvolution over groups. *IEEE Trans. Inform. Theory*, 54(1) :289–298.

[Kurtzman, 2009] Kurtzman, A. (2009). The ode method for some self-interacting diffusions. *Ann. Inst. H. Poincaré Probab. Statist., to appear.*

[Kushner and Yin, 2003] Kushner, H. J. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition. Stochastic Modelling and Applied Probability.

[Kusuoka and Stroock, 1987] Kusuoka, S. and Stroock, D. (1987). Applications of the Malliavin calculus. III. *J. Fac. Sci. Univ. Tokyo Sect. IA Math.*, 34(2) :391–442.

[Lagnoux, 2006] Lagnoux, A. (2006). Rare event simulation. *Probab. Engrg. Inform. Sci.*, 20(1) :45–66.

[Lagnoux-Renaudie, 2009] Lagnoux-Renaudie, A. (2009). A two-step branching splitting model under cost constraint for rare event analysis. *J. Appl. Probab.*, 46(2) :429–452.

[Le, 1998] Le, H. (1998). On the consitency of procrustean mean shapes. *Advances in Applied Probability*, 30 :53–63.

[Le and Kume, 2000] Le, H. and Kume, A. (2000). The fréchet mean shape and the shape of the means. *Advances in Applied Probability*, 32 :101–113.

[Le Cam, 1973] Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1 :38–53.

[Le Cam, 1986] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York.

[Lepski, 1991] Lepski, O. (1991). Asymptotically minimax adaptive estimation i. upper bounds, optimally adaptive estimates. *Theory Probab. Appl.*, 36(3) :682 ?697.

[Liu and Muller, 2004] Liu, X. and Muller, H. (2004). Functional convex averaging and synchronization for time-warped random curves. *Journal of the American Statistical Association*, 99(467) :687–699.

[Lutz and Bühlmann, 2006] Lutz, R. W. and Bühlmann, P. (2006). Boosting for high-multivariate responses in high-dimensional linear regression. *Statist. Sinica*, 16(2) :471–494.

[Meinshausen et al., 2007] Meinshausen, N., Rocha, G., and Yu, B. (2007). A tale of three cousins : Lasso, L_2Boosting and Dantzig. *Ann. Statist.*, 35(6) :2373–2384.

[Miclo, 1992] Miclo, L. (1992). Recuit simulé sur $\mathbf{R}^n$. Étude de l'évolution de l'énergie libre. *Ann. Inst. H. Poincaré Probab. Statist.*, 28(2) :235–266.

[Miles, 1982] Miles, J. W. (1982). On a nonlinear Bessel equation. *SIAM J. Appl. Math.*, 42(1) :109–112.

[Miller and Younes, 2001] Miller, M. I. and Younes, L. (2001). Group actions, homeomorphisms, and matching : A general framework. *International Journal of Computer Vision*, 41 :61–84.

[Newman, 2006] Newman, M. (2006). Finding community structure in networks using the eigenvectors of matrices. *Pysical Review E*, (036104) :709–718.

[Oyet and Wiens, 2000] Oyet, A. J. and Wiens, D. P. (2000). Robust designs for wavelet approximations of regression models. *J. Nonparametr. Statist.*, 12(6) :837–859.

[Pascucci and Polidoro, 2006] Pascucci, A. and Polidoro, S. (2006). Harnack inequalities and Gaussian estimates for a class of hypoelliptic operators. *Trans. Amer. Math. Soc.*, 358(11) :4873–4893 (electronic).

[Pemantle, 1992] Pemantle, R. (1992). Vertex-reinforced random walk. *Probab. Theory Related Fields*, 1 :117–136.

[Polidoro, 1997] Polidoro, S. (1997). A global lower bound for the fundamental solution of Kolmogorov-Fokker-Planck equations. *Arch. Rational Mech. Anal.*, 137(4) :321–340.

[Polyak, 1987] Polyak, B. (1987). *Introduction to Optimization.* Optimization Software, New York.

[Pronzato, 2000] Pronzato, L. (2000). Adaptive optimization and D-optimum experimental design. *Ann. Statist.*, 28(6) :1743–1761.

[Ramsay and Li, 2001] Ramsay, J. and Li, X. (2001). Curve registration. *Journal of the Royal Statistical Society (B)*, 63 :243–259.

[Rassmussen, 1994] Rassmussen, P. (1994). The pot method for flood estimation : a review. *Stoc. and Stat. Meth. in Hydro. and Environ. Eng.*

[Ratkowsky, 1983] Ratkowsky, D. (1983). *Nonlinear regression modeling.* Marcek Dekker Inc.

[Reynaud-Bourret, 2003] Reynaud-Bourret, P. (2003). Adaptive estimation of the intensity of inhomogeneous poisson processes via concentration inequalities. *Probability Theory and Related Fields*, 126 :103–153.

[Risken, 1989] Risken, H. (1989). *The Fokker-Planck equation*, volume 18 of *Springer Series in Synergetics*. Springer-Verlag, Berlin, second edition. Methods of solution and applications.

[Rossi and Villa, 2010] Rossi, F. and Villa, N. (2010). Optimizing an organized modularity measure for topographic graph clustering : a deterministic annealing approach. *Neurocomputing*, (73) :1142–1163.

[Rousseau, 2010] Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.*, 38(1) :146–180.

[Royer, 1989] Royer, G. (1989). A remark on simulated annealing of diffusion processes. *SIAM J. Control Optim.*, 27(6) :1403–1408.

[Samaria et al., 1994] Samaria, F. S., Samaria, F. S., Harter, A., and Site, O. A. (1994). Parameterisation of a stochastic model for human face identification.

[Sansonnet, 2011] Sansonnet, L. . (2011). Wavelet thresholding estimation in a poissonian interactions model with application to genomic data. *available at http ://arxiv.org/abs/1107.4219*.

[Sheu, 1986] Sheu, S. J. (1986). Asymptotic behavior of the invariant density of a diffusion Markov process with small diffusion. *SIAM J. Math. Anal.*, 17(2) :451–460.

[Speckman, 1985] Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.*, 13(3) :970–983.

[Stroock and Varadhan, 1972] Stroock, D. W. and Varadhan, S. R. S. (1972). On the support of diffusion processes with applications to the strong maximum principle. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. III : Probability theory*, pages 333–359, Berkeley, Calif. Univ. California Press.

[Temlyakov and Zheltov, 2011] Temlyakov, V. N. and Zheltov, P. (2011). On performance of greedy algorithms. *J. Approx. Theory*, 163(9) :1134–1145.

[Tikhonov, 1943] Tikhonov, A. N. (1943). On the stability of inverse problems. *C. R. (Doklady) Acad. Sci. URSS (N.S.)*, 39 :176–179.

[Trélat, 2005] Trélat, E. (2005). *Contrôle optimal*. Mathématiques Concrètes. [Concrete Mathematics]. Vuibert, Paris. Théorie & applications. [Theory and applications].

[Trèves, 1980] Trèves, F. (1980). *Introduction to pseudodifferential and Fourier integral operators. Vol. 1*. Plenum Press, New York. Pseudodifferential operators, The University Series in Mathematics.

[Tropp, 2004] Tropp, J. A. (2004). Greed is good : algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10) :2231–2242.

[Trouvé and Younes, 2005] Trouvé, A. and Younes, L. (2005). Metamorphoses through lie group action. *Foundations of Computational Mathematics*, 5(2) :173–198.

[Tsybakov, 2003] Tsybakov, A. (2003). *Introduction à l'estimation non-paramétrique*. Springer-Verlag, Paris.

[Tsybakov, 2004] Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1) :135–166.

[van de Geer, 2008] van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2) :614–645.

[van de Geer and Bühlmann, 2009] van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3 :1360–1392.

[van der Vaart and Wellner, 1996] van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes.* Springer Verlag, New York.

[Van der Waart, 1998] Van der Waart, A. (1998). *Asymptotic statistics*, volume 27 of *Cambridge Series in Statistical and Probabilistic Mathematics 03.* Cambridge Univ. Press, New York.

[Villani, 2006] Villani, C. (2006). Hypocoercive diffusion operators. In *International Congress of Mathematicians. Vol. III*, pages 473–498. Eur. Math. Soc., Zürich.

[Villani, 2009] Villani, C. (2009). Hypocoercivity. *Mem. Amer. Math. Soc.*, 202(950) :iv+141.

[Vimond, 2010] Vimond, M. (2010). Efficient estimation for a subclass of shape invariant models. *Annals of statistics*, 38(3) :1885–1912.

[Wahba, 1990] Wahba, G. (1990). *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

[Wang and Gasser, 1997] Wang, K. and Gasser, T. (1997). Alignment of curves by dynamic time warping. *Annals of Statistics*, 25(3) :1251–1276.

[Wolpert et al., 2003] Wolpert, R. L., Ickstadt, K., and Hansen, M. B. (2003). A nonparametric Bayesian approach to inverse problems. In *Bayesian statistics, 7 (Tenerife, 2002)*, pages 403–417. Oxford Univ. Press, New York. With a discussion by Subhashis Ghosal and a reply by the authors.

[Yazici, 2004] Yazici, B. (2004). Stochastic deconvolution over groups. *IEEE Trans. Inform. Theory*, 50(3) :494–510.

[Younes, 2004] Younes, L. (2004). *Invariance, déformations et reconnaissance de formes*, volume 44 of *Mathématiques & Applications (Berlin) [Mathematics & Applications].* Springer-Verlag, Berlin.

[Zhu et al., 2003] Zhu, J., Rosset, S., Hastie, T., and Tibshirani., R. (2003). 1-norm support vector machines. *In Proceedings of the 16th 2003 Conference Advances in Neural Information Processing Systems.*

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2) :301–320.