Document de synthèse présenté pour l'obtention d'une

Habilitation à Diriger des Recherches

Problèmes statistiques en grande dimension: sélection de variables et inférence de modèles déformables. Optimisation par dynamique non réversible.

présenté par

Sébastien Gadat

au vu des rapports de

Pierre Del Moral	Directeur de Recherche, Inria Bordeaux
Gabor Lugosi	Professeur, Université Pompeu Fabra
Stéphane Mallat	Professeur, École Polytechnique

soutenue le 22 novembre 2012, devant le jury composé de

Patrick Cattiaux	Professeur, Université Paul Sabatier	(coordinateur)
Djalil Chafaï	Professeur, Université Paris-Est	(examinateur)
Pierre Del Moral	Directeur de Recherche, Inria Bordeaux	(rapporteur)
Fabrice Gamboa	Professeur, Université Paul Sabatier	(examinateur)
Stéphane Mallat	Professeur, École Polytechnique	(rapporteur)
Clémentine Prieur	Professeur, Université Joseph Fourier	(examinateur)

Institut Mathématiques de Toulouse
Université Paul Sabatier

Remerciements

Je remercie sincèrement Pierre Del Moral, Gabor Lugosi et Stéphane Mallat d'avoir accepté de rapporter ce mémoire et de leur intérêt pour mon travail malgré tout le temps que cela demande. C'est pour moi un grand honneur de compter des scientifiques aussi remarquables dans mon jury et tient à leur exprimer toute ma reconnaissance. Je n'oublie bien sûr pas Patrick Cattiaux qui a coordonné mon habilitation avec bienveillance, ainsi que Djalil Chafaï, Fabrice Gamboa et Clémentine Prieur qui ont accepté de prendre part au jury.

Mes pensées vont ensuite à Laurent Younes qui m'a mis le pied à l'étrier durant ma thèse il y a quelques années et à qui je dois mon goût pour les mathématiques appliquées, ainsi qu'à Robert Azencott qui m'a donné mon premier cours de probabilités. De même, je voudrais exprimer tous mes remerciements à Alain Trouvé qui a toujours fait part de son intérêt pour les travaux que j'ai menés récemment.

Il y a des personnes avec qui j'ai développé d'intenses projets de recherche, et assurément sans qui l'envie d'aller à l'IMT aurait été bien moindre tous les matins. Fabien, Jérémie, j'espère réellement continuer à penser et travailler à vos côtés pendant longtemps ! Je tiens à souligner également l'immense plaisir que j'ai à partager de manière récurrente mes trop nombreuses questions aléatoires avec Laurent Miclo dont la finesse et l'originalité de son analyse probabiliste me laissent souvent pantois (sauf quand il se met à discuter de statistiques !). Je remercie également pour leur sérieux tous mes co-auteurs qui n'ont pas encore été évoqués : Jean-Marc Azais, Dominique Bontemps, Alexandre Cabot, Christine Cierco, Serge Cohen, Sébastien Déjean, Hans Engler, Thierry Klein, Agnès Lagnoux, Clément Marteau, Clément Pellegrini, Kim-Anh Lê Cao, Matthieu Vignes, Nathalie Villa. Je n'oublie également pas les chercheurs de l'équipe "Mathématiques pour l'Industrie et la Physique" avec qui j'ai pu discuter sur des problèmes de contrôle et optimisation, parmi eux Sylvain Ervedoza, Jean-Baptiste Hirriart-Urruty et Jean-Michel Roquejoffre, ainsi que les Tourangeaux Guy Barles et Emmanuel Lesigne.

Pour les nombreuses discussions scientifiques ou réflexions sur toutes autres choses, j'ai toujours apprécié les moments passés en la compagnie de Jean-Marc Azais, Dominique Bakry, Philippe Berthet, Jérôme Bertrand, Patrick Cattiaux, Fabrice Gamboa, Arnaud Guillin, Aldéric Joulin, Nicolas Savy et tant d'autres que j'oublie certainement. Je souhaite également un épanouissement complet à Aurélien, Xavier, Sébastien, Cathy, Pierre, aux thésards qui n'en sont déjà plus (ou presque plus) : Rim, Emmanuel, Yohann, Thibault, Thibaut, Paul ainsi qu'à Magali et Claire qui essuient les plâtres en m'expérimentant comme directeur...

Enfin, un grand merci à Cécile pour son accueil chaleureux à Lyon pendant deux mois d'hiver !

La vie d'un enseignant chercheur comportant également une seconde facette du métier, parfois salvatrice, je ne manquerais pas d'adresser tous mes remerciements à Eric Lombardi, André Legrand et Philippe Berthet pour m'avoir fait confiance en me laissant les clefs de la prépa agreg pendant ces très belles années écoulées. Je remercie bien sûr Patrick Martinez et Matthieu Hillairet de m'avoir aménagé une transition douce lors de cette prise de fonction, et Stéphane Lamy et Marc Perret pour m'avoir restitué ma liberté cette année ! J'ai également une

pensée pour tous les enseignants avec qui j'ai partagé cette expérience, ainsi que pour Sylvie Crabos qui m'a toujours accueilli avec toute sa bonne humeur dans son secrétariat. Enfin, je souhaite bon vent à tous les étudiants passés par cette prépa-agreg et les remercie pour toutes les émotions, qu'elles soient positives ou hélas parfois négatives, qu'ils m'ont fait partager.

Bien entendu, je serais totalement ingrat si je n'avais un mot pour Marie-Laure Ausset, Dominique Barrère, Delphine Dalla-Riva, Marie-Line Domenjole et Francoise Michel qui permettent au jour le jour de transformer tout problème en solution (il faudra un jour qu'elles me donnent le truc, je pourrais peut être ainsi alléger mes journées !).

J'aurais également tant à dire à mes parents, famille, amis, à vous Mélanie, Alban et Cerise ; mais qu'il me soit permis de vous remercier de votre soutien constant ailleurs qu'au travers de ces quelques lignes.

Table des matières

Chapitre 1

Introduction

Dans ce premier chapitre, je présente les thèmes de recherche abordés durant les 6 ans qui ont suivi la fin de ma thèse. Tous les thèmes sont mis en perspectives par rapport aux travaux qu'on peut trouver dans l'état de l'art. Certains de ces thèmes ne sont mentionnés que dans cette partie introductive tandis que certaines contributions de mes recherches sont par ailleurs plus détaillées dans les chapitres suivants (2, 3 et 4) ainsi que des élargissements possibles.

1.1 Statistiques en grande dimension

L'étude de problèmes d'estimation en grandes dimensions est un des enjeux majeurs des problèmes statistiques actuels. Étant données des observations $(X_1, \ldots, X_n)$ « étiquettées », une question importante est de savoir déduire une règle de prédiction de l'étiquette Y connaissant une nouvelle valeur des observations X sans bien entendu connaître la loi du couple. S'il existe désormais des techniques efficaces et classiques lorsque X appartient à un espace de petite dimension, le problème est tout autre dans le cas inverse. L'émergence d'un tel thème pour les statistiques mathématiques provient principalement de l'appartition de problèmes concrets de traitement du signal, d'images ou de bases de données. Ces problèmes peuvent être multiples et il serait vain d'essayer de tous les lister exhaustivement. Par contre, ils présentent tous un point commun qui est de devoir faire face à la malédiction des grandes dimensions : chercher à apprendre à partir d'un échantillon de taille n, un objet dont la structure est par nature dans un espace de dimension p avec p trop grand par rapport à n.

1.1.1 État de l'art

Cette problématique est très largement étudiée depuis une quinzaine d'années dans le cadre des problèmes de régression où les variables aléatoires Y sont réelles. Différentes approches ont été développées pour répondre à la question de sa prédiction par le biais de fonctions de X lorsque les données d'entrée sont de taille p très grande par rapport à n. Il est possible de dénombrer différentes familles de méthodes, chacune ayant pour objectif de construire $\hat{f}_{n,p}$ minimisant une fonction de perte L en général quadratique

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

où l'espérance est prise sur la loi *inconnue* des observations (X, Y). On peut mentionner rapidement différentes idées qui permettent de contourner le problème de l'estimation en grande dimension.

Méthodes pénalisées Ces méthodes permettent de procéder à une estimation de f sans forcément parvenir à opérer une sélection de variables. C'est le cas lorsqu'on cherche à estimer f par le biais d'un modèle linéaire $f(X) = {}^t\theta X$ en pénalisant la perte par la norme L^2 des coefficients de régression. Ainsi, on cherche à minimiser

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

où $\|\|_n$ désigne la norme empirique. Lorsque $p_n(\theta) = \lambda_n\|\theta\|_2^2$, on obtient la régression *Ridge* introduite dans [Hoerl and Kennard, 1975] et basée sur un procédé de Tikhonov [Tikhonov, 1943] qui consiste à régulariser un problème inverse mal posé, ce qui est précisément le cas lorsque $p >> n$ avec un modèle linéaire. Cette remarque a historiquement ensuite permis de construire des estimateurs basés sur des prédicteurs plus riches que des prédicteurs linéaires comme les splines dans des espace de Hilbert à noyau reproduisant (décrits dans [Wahba, 1990] par exemple). Il est important de noter que la présence et calibration de la pénalisation biaise nécessairement l'estimation même s'il est possible de donner des solutions pour la rendre évanescente lorsque n devient grand (critères AIC et BIC donnés dans [Akaike, 1974] ou plus récemment dans [Barron et al., 1999] par exemple).

Méthodes algorithmiques On peut également citer d'autres méthodes susceptibles de modérer la tendance à l' *overfitting* dans des contextes de grandes dimensions. Ces méthodes sont basées sur des idées plus algorithmiques comme la régression par algorithme CART [Breiman et al., 1984] ou Random Forests [Breiman, 2001, Amit and Geman, 1997] mais n'opèrent pas naturellement de sélection de variables. Dans la méthode CART, c'est un critère d'arrêt jouant le rôle de pénalisation qui permet de se prémunir du piège de la grande dimension. Dans les algorithmes Random Forests, c'est la randomisation et l'agrégation de prédicteurs décorrélés (amélioration du bagging de Breiman) qui permet de supprimer la tendance au sur-apprentissage lorsque $n << p$, comme démontré récemment par [Biau et al., 2008] Cette idée d'agrégation d'estimateurs est par ailleurs largement développée dans certains travaux de Tsybakov pour des contextes de classification notamment ([Tsybakov, 2004] par exemple).

Analyse multi-résolution Dans la situation où f est décrite par une famille dénombrable de coefficients, certaines méthodes supposent que la fonction f cible appartient à un espace ayant des propriétés de régularité (en général des espaces de Sobolev ou Besov) pour effectuer alors une analyse multi-résolution. Les travaux [Donoho et al., 1995, Donoho and Johnstone, 1995] proposent des méthodes de seuillage dans des bases d'ondelettes qui permettent à la fois de limiter le nombre de coefficients d'ondelettes et maintenir des propriétés de reconstruction qui peuvent être adaptatives en la régularité s de la fonction f à estimer. Par ailleurs, il est possible de rendre ces méthodes de reconstruction optimales au sens minimax pour la perte quadratique [Donoho and Johnstone, 1998]. C'est donc l'hypothèse d'espace fonctionnel qui rend alors le problème résoluble "statistiquement".

Méthodes *sparses* Depuis 10 ans, de nombreux travaux concernant l'estimation de f linéaire en les observations X a connu un tournant majeur au travers de la découverte de la méthode *Non-Negative Garotte* de [Breiman, 1995] qui a inspiré ensuite la méthode Lasso [Efron et al., 2004]. L'idée fondamentale à la base du Lasso consiste à exploiter la géométrie de la boule unité ℓ^1 et remarquer que les solutions du problème (1.1) lorsque $p_n(\theta) \propto \|\theta\|_{\ell^1}$ vont donner des solutions creuses. Ainsi, la résolution de cette optimisation aboutit implicitement à une méthode de sélection de variables qui controle le sur-apprentissage de l'estimateur. De très nombreux travaux exploitent et généralisent cette idée parmi lesquels l'Elastic Net de [Zou and Hastie, 2005],

le Dantzig Selector de [Candes and Tao, 2007] et de multiples travaux sur le Lasso (une liste loin d'être exhaustive étant par exemple [van de Geer and Bühlmann, 2009, van de Geer, 2008, Bickel et al., 2009]). La consistance statistique de telles méthodes est démontrée sous des hypothèses sur f. On suppose généralement que f est s parcimonieuse, et même si cette hypothèse n'est pas équivalente au cadre fonctionnel du paragraphe précédent, c'est une hypothèse de structure sur le signal à reconstruire qui permet l'estimation. Par ailleurs, la taille de l'échantillon n ne peut être arbitrairement petite devant p puisque les hypothèses nécessaires pour obtenir des résultats de consistance sont que $\log p$ et n sont du même ordre. À nouveau, ceci n'est pas sans rappeler les méthodes de seuillage dans les bases d'ondelettes utilisées en statistiques où le logarithme du nombre de coefficients gardés pour l'estimation est de l'ordre de n.

Algorithmes gloutons Enfin, il convient de mentionner des méthodes tirées de la théorie de l'approximation qui appartiennent à la famille des *Greedy Algorithms* initiés par les travaux de [DeVore and Temlyakov, 1996]. Ces méthodes itératives consiste dans la situation déterministe à construire à partir d'un dictionnaire (non nécessairement orthogonal) décrivant les données X des suites d'estimateurs de plus en plus précis de f. Dans la communauté statistique, ces algorithmes sont connus sous le nom d'algorithmes de *Boosting* tandis que la communauté d'approximation se réfère plutôt à des algorithms de *Matching Pursuit* de [Davis et al., 1994]. Là encore, une très large littérature existe sur ces algorithmes [Binev et al., 2005, Donoho et al., 2006, Donoho et al., 2007] qui mettent en jeu des propriétés de meilleures approximations par le biais d'inégalités de Lebesgue. L'idée principale de ces algorithmes est de construire séquentiellement des estimateurs diminuant à chaque étape la norme du résidu entre f et son approximation. Cette idée a alors été exploitée en statistiques dans les travaux de [Bühlmann and Yu, 2003] puisque ces algorithmes munis d'un critère d'arrêt adéquat permettent une représentation parcimonieuse de l'estimateur et possèdent des propriétés de bonne reconstruction asymptotique avec grande probabilité, même dans une situation où les prédicteurs sont corrélés entre eux. Remarquons enfin que des idées tout à fait similaires avaient déjà été développées dans la communauté d'apprentissage statistique pour des problèmes de classification ([Freund and Schapire, 1997]). L'idée est alors de fabriquer séquentiellement des prédictions se concentrant sur les individus les plus difficiles à classer / prédire. Tout comme pour l'algorithme [Bühlmann and Yu, 2003], le critère d'arrêt de ces algorithmes joue un rôle fondamental pour sa performance numérique.

1.1.2 Travaux de thèse - Sélection de variables et classification

Lors de ma thèse [1] éffectuée sour la direction de Laurent Younes, j'ai travaillé sur la thématique de la classification supervisée de données en grande dimension. Plus précisément, étant donnés un échantillon de n individus préalablement classés décrits par un grand nombre p de variables descriptives, la question posée était de réussir à sélectionner parmi les p variables un petit nombre d'indices pertinents. Le but était alors double : améliorer les performances de classification et en parallèle isoler dans la base de données ce qui structure le problème de classification. Le premier aspect est intéressant dans un contexte d'efficacité de classification alors que le second peut revêtir une grande importance, par exemple en biologie.

Si la litterature sur les problèmes de sélection de variables pour la régression est assez largement fournie désormais, il faut noter que c'est moins le cas dès lors que l'apprentissage concerne un problème de classification. En général, les algorithmes connus se distinguent en deux classes : les premiers sont des algorithmes de filtrage qui fonctionnent quelle que soit la nature de la méthode de classification $\mathbb{A}$ mise en place par la suite. Ils sont en général basés sur des heuristiques. On pourra consulter par exemple [Guyon et al., 2006] pour une liste

assez complète de toutes les méthodes filtres de sélection de variables. Cependant, la plupart de ces méthodes ne disent rien théoriquement sur leurs consistances statistiques. Les seconds algorithmes sont de type *wrappers*. Ils sont basés sur une optimisation récursive d'un critère dédié ou non à la méthode de classification $\mathbb{A}$ utilisée. On peut par exemple citer *Recursive Feature Elimination* [Guyon et al., 2002] qui supprime des variables ayant peu d'influence sur la marge de classification d'un SVM. On peut également relever les récents développements basés sur une modification de l'algorithme de *Support Vector Machine* auquel est adjoint une pénalisation ℓ^1 [Bi et al., 2003, Zhu et al., 2003], s'inspirant ainsi de l'algorithme Lasso.

L'algorithme développé dans ma thèse appartient à la seconde famille de méthodes, mais est un petit peu différent puisqu'il fonctionne quelle que soit la nature de $\mathbb{A}$. Plus précisément, si $\mathcal{D}$ désigne le dictionnaire de variables descriptives pour les individus $X_1, \dots X_n$ étiquettés par des variables qualitatives $(Y_1, \dots Y_n)$ et si l'on considère un algorithme de classification supervisée $\mathbb{A}$, l'objectif a été de développer une méthode de type « *best subset* » pour isoler un sous-ensemble $\omega \subset \mathcal{D}$. Le parcours de tous les sous-ensembles de $\mathcal{D}$ étant numériquement irréalistes, le travail a consisté à produire un algorithme stochastique effectuant un parcours non exhaustif des sous-ensembles de $\mathcal{D}$. L'objectif est d'exhiber une méthode d'optimisation de ω pour les performances de $\mathbb{A}$ en exploitant la base de données $(X_1, Y_1), \dots, (X_n, Y_n)$.

À la fin de mes travaux de thèse, deux articles théoriques ont été écrits sur le sujet. Le premier [6] décrit l'algorithme stochastique et propose des applications en traitement du signal lorsque l'algorithme de référence $\mathbb{A}$ est un algorithme de *Support Vector Machine*. L'idée de cet algorithme est de fonctionner comme une méta-méthode permettant d'optimiser un poids sur chacune des variables du dictionnaire $\mathcal{D}$. Son fonctionnement est itératif et évoque par conséquent un petit peu le mécanisme de [Bühlmann and Yu, 2003] et [Freund and Schapire, 1997] puisqu'on diminue d'une itération à l'autre le poids de certaines variables proportionnellement à l'erreur engendrée sur la base de données. Ainsi, c'est un algorithme d'apprentissage de boosting sur les variables descriptives.

Le second travail [5] généralise la méthode de parcours des sous-ensembles en proposant des règles de composition en arbres binaires des variables de $\mathcal{D}$. Il est autant d'inspiration algorithmique [Breiman, 2001] pour les parcours d'arbres que théoriques puisque la stratégie de parcours stochastique réversible dans des forêts d'arbres binaire était à construire. Dans ce manuscrit, j'ai choisi de ne rappeler que le modèle de sélection de variables ainsi que l'algorithme stochastique initial qui en a découlé dans le paragraphe 2.1 puisque certains développements ultèrieurs ont été motivés par ce travail.

1.1.3 Applications à des données bio-statistiques

Ma thèse ayant porté sur les problèmes d'estimation en grande dimension, j'ai naturellement été conduit à travailler avec les chercheurs du groupe « biopuce » qui précisément étudiaient des méthodes statistiques de classification supervisée pour les appliquer à des données de type *Microarrays* fournies par l'INRA. Avec Kim-Anh Lê Cao, nous avons cherché à étendre le champ de simulations à des données biopuces en utilisant comme algorithme $\mathbb{A}$ un algorithme CART de classification par arbre binaires [2] puis nous avons ensuite considéré un cadre multi-classes [3] en s'attachant à étudier tant les performances de classification que les stabilités des résultats obtenus et les interprétations biologiques qui en découlent. Même si l'on peut considérer que cette dernière collaboration s'est essentiellement cantonée à des questions d'applications numériques, elle a été extrêmement enrichissante pour différentes raisons décrites en particulier dans les paragraphes 1.2.3, 1.3 et 1.4.

1.2 Problèmes d'estimations en grande dimension

Je présente dans ce paragraphe une brève description des travaux développés depuis ma thèse et concernant des problèmes d'estimation statistiques en grande dimension. Tous ces problèmes possèdent un dénominateur commun qui est l'inférence d'événements « rares » à la vue du nombre d'expériences dont on dispose dans une base de données. Ces travaux sont tous d'inspiration algorithmique mais certains présentent des avancées théoriques (paragraphes 1.2.1 et 1.2.3), d'autres de modélisations (paragraphe 1.2.2) et enfin les derniers concernent des applications industrielles de certaines méthodes stochastiques (paragraphe 1.2.4).

1.2.1 Plans d'expériences séquentiels

Le premier travail se situe dans la thématique des plans d'expériences dont est issue une collaboration avec Serge Cohen et Sébastien Déjean. La problématique provient d'une question concernant les gros codes numériques. On peut modéliser ces codes comme une fonction f à interpoler en minimisant le nombre de points de mesure à utiliser pour calculer une estimation $\hat{f}$ de f. Lorsque l'estimateur construit est linéaire, il existe des manières relativement explicites de quantifier l'efficacité escomptée de l'estimation en terme de variance de $\hat{f}$. On pourra par exemple se référer aux premiers travaux de [Kiefer and Wolfowitz, 1959, Fedorov, 1972] qui étant donnée une famille de prédicteurs, donnent plusieurs critères d'optimalité de plans d'expérience pour des modèles linéaires. Le point de vue adopté a été de construire des plans séquentiels (on choisit de positionner le point x_{k+1} après avoir mesuré $f(x_k)$) à la manière des travaux développés dans [Pronzato, 2000] tout en laissant la possibilité ou non à l'utilisateur de contrôler le biais du modèle par une approche minimax [Oyet and Wiens, 2000]. Par ailleurs, nous proposons une approche où la famille des prédicteurs utilisés n'est pas figée mais peut varier stochastiquement au cours des itérations, ce qui permet une alternative à la méthode développée dans [Biswas and Chaudhuri, 2002] qui est uniquement basée sur une stratégie de tests *backward*.

En exploitant la stratégie stochastique de parcours d'arbres déjà utilisée dans [5], nous proposons dans [4] un nouvel algorithme stochastique basé sur une analyse multi-résolution pour construire récursivement les points de designs optimaux d'interpolation de f. En plus de cette nouvelle méthode stochastique de construction de plans d'expériences séquentiels, nous prouvons un théorème de localisation de designs optimaux dédié au cas particulier de la base multi-résolution de Schauder qui rend la méthode très rapide dans le cas d'utilisation d'une telle base pour estimer f. Ce résultat de localisation est non trivial puisqu'il est basé sur une famille de fonctions qui n'est pas un T-système (voir la définition des systèmes de Tchebychev dans [Dette and Studden, 1997]). Ce travail est présenté dans le paragraphe 2.2.

1.2.2 Reconstruction de graphes de communauté

En collaboration avec Nathalie Villa, nous developpons dans [7] un algorithme de clustering de graphes avec comme point de mire l'application à l'identification de partitions de sommets dans les graphes de communautés. Un graphe est défini au travers de sa matrice d'adjacence W codant la présence d'arêtes entre les différents sommets du graphe. L'idée a été d'exploiter un a priori sur les graphes de communautés qui est l'existence de sous-ensembles de sommets du graphe, fortement connectés entre eux et faiblement connectés aux autres sommets (ces sous-ensembles formant alors une communauté). Généralement, le clustering de graphe peut être résolu en utilisant la diagonalisation de la matrice d'adjacence [Newman, 2006]. Ceci étant, une telle méthode peut ne pas être adaptée par rapport à la structure de communauté attendue. Étant donné un graphe non orienté défini par W symétrique à diagonale nulle, on définit pour

chaque noeud i le nombre d'arêtes reliées à i (son degré). Plus précisément, $W_{i,j} = W_{j,i} = 1$ si i et j sont connectés tandis que $W_{i,j} = W_{j,i} = 0$ sinon et bien sûr $d_i = \sum W_{i,j}$. En notant $2m$ le nombre d'arêtes codées dans W, et pour une partition $C_1, \dots C_k$ des sommets, on définit la Q modularité par

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

On constate alors qu'une partition qui possède une grande Q-modularité est le signe d'une sur-représentation d'arêtes intra-classe. Le parcours de toutes les partitions étant infaisable pour des graphes de taille importante, nous avons alors opté dans [7] pour l'application d'un algorithme de recuit simulé afin d'optimiser cette fonction Q. À noter que le travail a consisté également à développer un algorithme de force pour la représentation des partitions obtenues après optimisation stochastique. Cette phase est aussi importante que la construction d'une bonne partition afin d'obtenir une visualisation lisible des résultats. On peut également noter un récent développement utilisant la même approche [Rossi and Villa, 2010] combinée à un recuit simulé déterministe approché par champ moyen.

1.2.3 Reconstruction de graphes de réseaux de gènes

Ce premier travail sur les apprentissages de structure de graphe et les contacts précédents tissés avec l'Inra m'ont conduit naturellement à interagir avec des chercheurs du laboratoire de Biométrie et Intelligence Artificielle. La reconstruction de réseaux de régulation de gènes peut être intéressante aussi bien pour des éclairages nouveaux sur les processus biologiques sous-jacents que pour des applications comme l'identification des mécanismes de maladies génétiques en vue d'un traitement ciblé.

Le contexte est le suivant : on mesure deux types de données sur un ensemble de n individus. Le premier type de données sont les données d'expression E de taille $n \times p$ où p désigne le nombres de gènes. Ce sont des variables quantitatives donnant un niveau d'expression de chaque gène sur chacun des individus. Les secondes données sont les données marqueurs M (de taille $n \times p$ également) qui sont des variables discrètes.

Dans le réseau, une interaction entre 2 gènes (*i.e.* le fait que la protéine issue d'un gène active ou inhibe l'expression de l'autre gène) est représentée par une arête entre ces gènes et l'interaction du réseau est modélisée au travers de la relation :

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

où ϵ est un l'écart entre l'interaction théorique du modèle et la réalité. Par ailleurs, β est une matrice à diagonale nulle qui code la structure du réseau de régulation. Les inconnues du problème sont donc les deux matrices α et β à re-construire et sa difficulté réside en le fait que le nombre d'individus n est en général petit devant p le nombre de gènes du réseau, et donc *a fortiori* des $2p^2 - p$ inconnues de (1.2).

Nous avons commencé par aborder dans [19] des méthodes de régression pénalisée d'un point de vue expérimental (régression Lasso, Elastic Net et Dantzig) avant d'opter pour une implémentation du *Boosting*, méthode ayant été étudiée dans [Lutz and Bühlmann, 2006] auparavant. Cependant, l'extensition au cas multivarié donnée dans [Lutz and Bühlmann, 2006] est plus guidée par une généralisation naturelle des preuves théoriques données dans [Bühlmann, 2006] et dont les points clefs sont décrits dans les premiers résultats de [DeVore and Temlyakov, 1996] dans un cadre déterministe. Finalement, [Lutz and Bühlmann, 2006] ne considère pas la répartition de l'« effort » multivarié de l'algorithme de *Boosting* en plongeant le problème dans un espace plus gros.

Dans [19], nous modifions donc l'algorithme de boosting pour qu'il s'adapte à la nature multivariée de la régression (1.2). Ainsi, par rapport aux travaux de [Bühlmann and Yu, 2003], nous avons introduit une étape supplémentaire de boosting pour choisir la coordonnée à prédire [1]. Cela nous a conduit à étudier d'abord le problème théorique déterministe ($\epsilon = 0$) en reprenant les preuves de [DeVore and Temlyakov, 1996] puis à étendre nos résultats à la situation plus réaliste bruitée. Ce travail est décrit dans le paragraphe 2.3.

1.2.4 Estimation par le biais de la théorie des valeurs extrêmes

Enfin, j'ai travaillé sur des problèmes très concrêts en collaboration contractuelle avec Thales Alenia Space et le Cnes de 2009 à 2011 et concernant le nouveau système de positionnement Galileo-Egnos. L'ESA exige de ce nouveau système qu'il fournisse un positionnement dans un tube de confiance imposé et que dans le cas contraire il retourne une alerte à l'utilisateur. Par ailleurs, la probabilité que le système ne retourne pas d'alerte à tort est imposée par l'ESA à $p = 10^{-7}$ sur une période de 150 secondes. Les vérités terrains ayant été recensées durant la période 2006-2009 ainsi que le positionnement donné par Egnos-Galileo, la question a été d'estimer la probabilité que le système ne retourne pas d'alerte à tort pour chaque station de mesure dispersée en Europe afin de certifier la valeur de p.

Bien entendu, des événements avec une si faible occurrence sont en pratique rarement observés, même sur un jeu de données de 3 ans. Procéder à une estimation de la probabilité de tels événements extrêmes ne peut se cantonner à effectuer une estimation de moyenne empirique. Le premier travail [26] a consisté à s'appuyer sur la théorie de valeurs extrêmes donnée par la loi de Fisher-Tippett (1928) qui sous réserve de certaines conditions d'indépendances des observations, décrit paramétriquement la loi des grandes valeurs d'un échantillon. Plus précisément, c'est l'aménagement par la méthode Peak Over Threshold (POT) (voir [Rassmussen, 1994, de Haan and Ferreira, 2006]) que nous avons utilisé pour effectuer les estimations de ces probabilités. Cette collaboration industrielle avec Cécile Mercadier et Jean-Marc Azaïs a donné lieu à un rapport technique [38] ainsi qu'à la constitution d'un premier logiciel d'estimation.

Un second travail a abordé le problème d'une manière relativement différente en utilisant des techniques de renforcement d'événements rares par algorithmes de Splitting en utilisant l'approche décrite dans [Lagnoux-Renaudie, 2009, Lagnoux, 2006]. L'idée ici est de procéder à des simulations hiérarchiques d'événements redoutés amenant alors à l'événement dont on veut estimer l'occurrence. À nouveau, une note technique [37] a été rédigée par Agnès Lagnoux, Cécile Mercadier et moi même pour Thales Alenia Space afin qu'ils comparent les résultats numériques avec des méthodes d'estimation basées sur les réseaux de Petri.

Un dernier travail sur ce même thème a consisté à reprendre l'étude utilisant la théorie des valeurs extrêmes pour automatiser les procédures statistiques utilisées dans [38]. Ces travaux sont intimement liés à l'estimation *a priori* d'un paramètre de seuillage des grandes observations d'un échantillon et donc à la question : « À partir de quand commencent les grandes valeurs de mon échantillon ? » Pour répondre à cette question, plusieurs algorithmes ont été employés notamment ceux de [Drees and Kaufmann, 1998, Beirlant et al., 1999] mais celui qui a retenu notre attention est décrit dans [de Sousa and Michailidis, 2004] et se base sur des résultats de distributions de sommes cumulées de grandes valeurs d'échantillons. Différentes difficultés techniques avaient été rencontrées, en particulier lors de la présence de corrélations dans les séries temporelles fournies par Thales, séries qui présentaient des phénomènes de non-stationnarité. Il faut noter qu'à partir de la dernière note technique [36] rédigée conjointement avec Jean-Marc

1. Ce qui en fait finalement un algorithme « boost-boost »

Azaïs, un processus de codage pour industrialisation a été lancé afin que l'outil algorithmique soit implanté sur le système Egnos-Galileo.

1.3 Traitement du signal et déformation

Le travail sur les données bio-puces m'a amené à la réflexion que la compréhension de la structure générative des données peut apporter plus comme information qu'une force brute algorithmique, si puissant soit-il pour résoudre un problème. Lors de ma thèse, j'avais travaillé sur les bases de données Mnist et US Postal de chiffres manuscrits qui sont une parfaite illustration du fait que certaines données manipulées peuvent être issues d'un processus de déformation aléatoire qu'il est parfois avantageux de modéliser le plus fidèlement possible afin d'estimer les paramètres génératifs pour un meilleur traitement des données.

1.3.1 État de l'art

Présentation des modèles déformables Jérémie Bigot venait d'étudier des méthodes statistiques de mise en correspondance de *landmarks* dans des images bruitées. Nous avons cherché à modéliser statistiquement le processus génératif des données en utilisant l'approche développée dans les travaux d'Alain Trouvé et Laurent Younes qui décrivaient principalement les situations déterministes. Nous avons principalement développé des techniques statistiques basées sur l'approche de Grenander [Grenander, 1993a, Grenander and Miller, 2007] pour modéliser l'espace des formes tandis que les déformations sont introduites par le biais des modèles de difféomorphismes [Younes, 2004].

En toute généralité, le modèle de déformation que nous étudions est décrit ainsi : si $f^\star$ est une forme de référence définie sur $\Omega \subset \mathbb{R}^d$, on observe des données

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

Les f_i correspondent alors à la forme $f^\star$ de référence et sont déformées aléatoirement tandis que les W_i représentent un bruit additif de mesure. Par ailleurs, les déformations possibles sont des éléments d'un groupe G de difféomorphismes de Ω. Ainsi,

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

où bien sûr $x \mapsto g_i.x$ désigne l'action de g_i sur le domaine Ω. Il se dégage alors deux familles de problèmes. Le premier se rapporte à l'estimation des g_i paramètres de déformations, le second correspond à l'estimation de $f^\star$. C'est plus sur ce second problème qu'ont porté mes travaux. Dans la plupart des situations, G est un groupe de Lie de dimension finie (déformations rigides) ou infinie (déformations élastiques). Notons qu'une simple moyenne empirique ne prenant pas en compte les effets de déformation ne peut être satisfaisante comme le montre la figure 1.1.

Ce phénomène de *blurring* illustré par la figure 1.1 montre que traiter l'estimation de $f^\star$ comme si les observations étaient dans un espace plat (paramétré par une distance euclidienne) est impossible. Plus précisément, étant donné un espace de Hilbert $\mathcal{H}$ qui contient les réalisations Y_i, la moyenne empirique est définie par le biais de la solution du problème de minimisation

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

Lorsque $\mathcal{H}$ est muni d'une distance euclidienne et que les déformations $g \in G$ sont de loi h, on observe alors par la loi des grands nombres que $\bar{Y}_n$ estime $\tilde{f}$ définie par

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

FIGURE 1.1 – Moyenne empirique « naïve » de 5 images de visage issus de la base de données Olivetti [Samaria et al., 1994].

et en toute généralité, $\tilde{f} \neq f^\star$ puisque c'est la convolution de $f^\star$ avec la loi des déformations (d'où le *blurring* observé pour la moyenne empirique).

Il est donc tentant de relier les estimations données par la définition (1.4) à des métriques différentes sur $\mathcal{H}$, notamment celles prenant en compte les effets de déformations. Cette approche est proposée dans les travaux de [Joshi et al., 2004, Miller and Younes, 2001, Trouvé and Younes, 2005] qui munissent alors $\mathcal{H} = L^2(\Omega)$ de la distance définie par

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

où e désigne l'élément neutre de G, λ un paramètre de pénalisation strictement positif, et D une mesure de distorsion entre g et l'absence de déformation.

L'utilisation de métriques différentes de la métrique euclidienne comme celle donnée par (1.5) pour l'estimation de $f^\star$ aboutit naturellement à la moyenne de Fréchet intrinsèque ou extrinsèque [Fréchet, 1948] des observations, notion généralisant la moyenne euclidienne dans des espaces courbés. Les comportements statistiques des estimateurs donnés par (1.4) dans le cas où les observations appartiennent à des variétés Riemanniennes de dimension finie sont assez bien connues. On pourra par exemple consulter [Bhattacharya and Patrangenaru, 2003, Bhattacharya and Patrangenaru, 2005] pour des résultats de consistance des moyennes de Fréchet extrinsèques vers la moyenne de Fréchet intrinsèque lorsque le nombre d'observations $n \mapsto +\infty$. Notons que ces résultats mélant loi des grands nombres (M-estimation) et géométrie riemanienne ont abouti dans [Le, 1998, Le and Kume, 2000] à des résultats de consistance des moyennes de Fréchet pour des courbes plannaires aléatoires appartenant à un cas particulier qui est l'espace des formes de Kendall défini dans [Kendall, 1984].

Enfin, ces travaux sont d'inspiration plutôt géométrique et ne s'intéressent pas tous aux problèmes d'estimations non-paramétriques naturels alors que les signaux observés sont des courbes ou des images.

Travaux en statistiques non-paramétriques Le problème de l'estimation de $f^\star$ sous jacent au modèle (1.3) a finalement été peu étudié dans le contexte des statistiques non paramétriques. On peut néanmoins citer les travaux précurseurs [Kneip and Gasser, 1988] qui introduisent le *shape invariant model* (modèle SIM) et proposent une estimation de $f^\star$ dans le cadre extrêmement simplifié où $d = 1$ (les observations sont des courbes paramétrisées par un nombre fini de coefficients) et où l'action de g est une translation de $\Omega = \mathbb{R}$:

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

D'autres travaux d'alignements de courbes ont été proposés par [Gasser and Kneip, 1992, Gasser and Kneip, 1995] dans un cadre semi-paramétrique et leur méthode est très fortement reliée aux moyennes de Fréchet précédemment évoquées. Les résultats font alors en général intervenir à la fois le nombre de courbes observées ainsi que le nombre de points d'observations par courbes. De même, les méthodes proposées dans [Wang and Gasser, 1997, Ramsay and Li, 2001, Liu and Muller, 2004] s'intéressent à des problèmes de déformations non nécessairement réduites aux translations mais ces travaux portent surtout sur la paramétrisation des bijections non rigides plutôt que sur des reconstructions asytmptotiques en n pour un cadre non-paramétrique.

Dans le contexte d'estimation des paramètres de déformations, [Gamboa et al., 2007a] ainsi que [Vimond, 2010] proposent des méthodes pour calculer des estimateurs des translations, et ensuite en déduire des reconstructions de $f^\star$. Signalons que [Bigot et al., 2010] généralise l'estimation des paramètres de déformations pour des groupes de Lie compacts paramétrant des déformations rigides non réduites aux translations.

Enfin, une approche récente développée par [Allassonière et al., 2007] propose un point de vue bayésien pour effectuer des estimations de $f^\star$ à partir des observations Y_i tandis que [Allassonière et al., 2009] implémente un algorithme stochastique (SAEM) afin de trouver l'optimiseur de la vraisemblance profilée des observations.

Dans tous les travaux statistiques précédents, la plupart ne propose de calculer des vitesses de reconstruction et parfois même la nature exacte de la limite des estimations proposées est inconnue ([Allassonière et al., 2007, Allassonière et al., 2009]). C'est donc dans ce contexte que je me suis intéressé au développement de méthodes fournissant des vitesses de reconstruction de $f^\star$.

1.3.2 Estimation de courbes décalées aléatoirement

Pour les signaux déformés aléatoirement, le modèle le plus simple auquel on peut penser (et pour lequel on peut analyser le plus finement les choses) est le suivant : on dispose d'une famille d'observations qui sont des courbes $(Y_i)_{i\in[1;n]}$ définies au travers du modèle de bruit blanc :

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

où $f^\star$ est la fonction à estimer est 1 périodique et définie de $\mathbb{R}$ dans $\mathbb{R}$. Le niveau de bruit est donné au travers de σ, $(W_i)_{i\in1\dots n}$ sont n mouvements Browniens et les $(\tau_i)_{i\in1\dots n}$ sont les n décalages aléatoires et modélisent ici le processus de déformation. En supposant les $(\tau_i)_{i\in1\dots n}$ i.i.d. et indépendants des $(W_i)_{i\in1\dots n}$, on cherche à donner une estimation de $f^\star$ et comprendre quels sont les mécanismes qui rendent le procédé d'estimation facile ou au contraire difficile.

Réponse asymptotique ($n \mapsto +\infty$) Nous donnons une première réponse à l'estimation de $f^\star$ pour le problème (1.6) pour un cadre asymptotique dans [9] en étudiant un estimateur de $f^\star$ basé sur un seuillage dur dans la base d'ondelettes de Meyer. L'obtention de la consistance ainsi que la vitesse de reconstruction comporte un certain nombre de points communs avec ce qui est fait en problèmes inverses statistiques résolus par déconvolution [Johnstone et al., 2004, Carroll and Hall, 1988] même si la présence d'un aléa supplémentaire sur les shifts τ_i engendre un surcroît de difficultés techniques pour les méthodes de seuillages des coefficients.

Par ailleurs, il est possible de calculer la vitesse minimax de l'estimation pour la norme L^2 lorsque $f^\star$ appartient à une boule de Besov $B^s_{p,q}(A)$. Le résultat le plus frappant est qu'alors le problème de reconstruction est relié à une méthode de déconvolution même si chaque courbe est une réalisation du modèle de bruit blanc shifté mais *non convolé*. À noter que l'obtention de la borne inférieure est basée sur une adaptation du Lemme d'Assouad décrit dans

[Bretagnolle and Huber, 1979, Has′minskiĭ and Ibragimov, 1990]. Cependant, l'idée défendue dans [Birgé, 1986] que l'approche des minorations par le lemme d'Assouad peut toujours se résoudre en utilisant le lemme de Fano semble également vraie ici même si l'aménagement de ce lemme de Fano (voir par exemple [Ibragimov and Has′minskiĭ, 1981]) revient en fin de compte aux mêmes calculs que ceux effectués en utilisant l'approche par lemme d'Assouad.

Inégalité oracle à horizon fini Dans [12], nous donnons une réponse non asymptotique en proposant un estimateur relié à $f^\star$ par le biais d'inégalités oracles. Ces travaux adaptent des techniques URE *Unbiased Risk Estimation* utilisées par exemple dans [Cavalier et al., 2002]. Dans notre contexte, on peut alors faire apparaître dans l'inégalité oracle à la fois un terme inhérent au modèle de bruit blanc et un terme faisant intervenir la complexité du problème inverse lorsqu'on utilise une déconvolution en base de Fourier et un filtre basse fréquence des coefficients (on pourrait également étendre à d'autres types de filtres comme ceux décrits par [Bissantz et al., 2007] dans notre contexte). On notera enfin que les objets utilisés dans [12] sont très proches des outils utilisés dans le cadre des problèmes inverses avec opérateurs inconnus partiellement observés [Cavalier and Raimondo, 2007, Cavalier and Hengartner, 2005].

Mes travaux sont tous basés ici sur une hypothèse discutable qui est la connaissance a priori de la loi des décalages $(\tau_i)_{i\in 1\ldots n}$. Il est néanmoins possible de proposer des estimations sortant de ce cadre en revenant alors aux moyennes de Fréchet [Bhattacharya and Patrangenaru, 2003], mais l'étude théorique est alors bien plus difficile puisque le contexte est ici non-paramétrique. Ces travaux sont décrits dans le paragraphe 3.2.1 du chapitre 3.

1.3.3 Estimation d'images par déformations rigides ou élastiques

Nous étendons le modèle de déformation aléatoire en enrichissant la structure des déformations agissant sur le signal $f^\star$ à reconstruire. Ainsi, il peut être naturel de considérer un groupe de déformations G plus gros que $(\mathbb{R}/\mathbb{Z}, +)$ qui est la situation décrite plus haut, notamment lorsque le signal $f^\star$ n'est plus une courbe mais une image et que G contient par exemple des translations ou des rotations. Une seconde étude asymptotique, plus générale que celle des courbes décalées, est proposée lorsque G est un groupe de Lie compact dans [8]. Les outils mis en jeu sont des éléments classiques de théorie des groupes de Lie (théorème de Peter-Weyl et transformée de Fourier) qui ont déjà été utilisés en statistiques dans un cadre de déconvolution « pur »par [Koo and Kim, 2008]. Ces travaux sont là-encore des généralisations de [Kim, 1998, Yazici, 2004]. La qualité d'estimation statistique repose à nouveau sur un contrôle fin dans le lemme d'Assouad des rapports de vraisemblance en fonction de la « taille » du groupe de Lie considéré.

Enfin, il est possible de modéliser des déformations plus complexes que des déformations rigides en générant de manière paramétrique des flots de champs de vecteurs, en utilisant principalement le modèle de grandes déformations de Trouvé et Younes. Nous exploitons ce modèle de grandes déformations paramétriques et des techniques de M-estimation (voir par exemple [Van der Waart, 1998]) pour étudier l'estimation d'une forme moyenne lorsque les données sont soumises à une déformation aléatoire bruitée. Des propriétés de consistance sont données dans [11]. Ces travaux font l'objet des paragraphes 3.3.5 et 3.2.2.

1.3.4 Régression sous contrainte de monotonie

De manière relativement anecdotique, signalons que la construction des flots de champ de vecteur nous a permis de donner des constructions simples de difféomorphismes, en par-

ticulier en dimension 1. Nous exploitons cette simple remarque pour construire des estimateurs dans des modèles de regression sous contrainte de monotonie de la fonction cible. Ce problème de régression sous contrainte de forme est l'objet de nombreux travaux en statistiques puisque parfois l'information a priori de monotonie de la prédiction est connue. Ainsi, [Hall and Huang, 2001, Dette et al., 2006, Dette and Pilz, 2006] abordent ces problèmes sous la forme d'une projection d'un estimateur obtenu par lissage splines sur un espace de fonctions monotones.

Nous avons pris le parti d'éviter l'opération de projection car elle introduit numériquement des artefacts. La méthode d'estimation donnée dans [10] exploite le fait que toutes les fonctions strictement croissantes de $[0;1]$ (par exemple) peuvent se décrire exactement comme les solutions au temps 1 de certaines équations différentielles basées sur des champs de vecteurs affines en temps. Ce travail est décrit dans le paragraphe 3.1.3.

1.3.5 Estimation d'intensité de processus de Poisson décalés aléatoirement

Suite aux travaux effectués dans le cas des translations aléatoires, j'ai été contacté par des chercheurs de l'Inserm de Toulouse afin de comprendre les déformations statistiques des processus de comptages de fixation de protéines le long de brins d'ADN. Le point remarquable est que ces processus de comptage donnant lieu aux données Chip-Seq sont souvent mal localisés : à savoir l'initialisation géographique des comptages est effectuée manuellement et peut être erronée. Cela amène le biologiste à effectuer une convolution (lissage) des données de comptage par noyau Gaussien afin d'obtenir une courbe continue puis effectuer un recalage et une moyenne de toutes ces données convolées pour en déduire des propriétés sur la fixation de certaines protéines dans différentes conditions expérimentales.

Il aurait été tentant[2] d'appliquer notre estimateur issu du modèle (1.6) à cet exemple de données. Néanmoins, la nature du processus étant relativement différente, nous avons proposé assez naturellement dans [17] une modélisation où chaque observation est la réalisation d'un processus de Poisson d'intensité inhomogène λ_i. Pour modéliser le problème d'initialisation du processus de comptage, tous les λ_i sont supposés être issus d'une intensité commune λ après une opération de shift aléatoire, ce modèle étant bien sûr inspiré de (1.6). À noter également que des problèmes un petit peu similaires [Sansonnet, 2011] peuvent apparaître en génétique lorsque les décalages sont observés ainsi que la moyenne des processus de comptage.

Une telle modélisation rend alors possible l'utilisation de propriétés de concentration pour des processus de Poisson [Reynaud-Bourret, 2003] et notre travail rentre alors dans le cadre des problèmes inverses poissonniens décrits par exemple dans [Cavalier and Koo, 2002, Kolaczyk, 1999] qui ont été abordés par le biais d'analyse multi-résolutions. Ainsi il est aussi possible de construire un estimateur asymptotiquement minimax par le biais d'une analyse en ondelettes des observations dans notre contexte.

Tout comme dans [9], une grande difficulté est la construction d'une borne inférieure faisant apparaître un problème inverse de déconvolution pour ce modèle statistique. Par ailleurs, proposer un estimateur adaptatif réclame également un effort non négligeable par rapport au modèle de bruit blanc puisqu'il est important d'estimer la norme $\|.\|_1$ de l'intensité du processus pour seuiller convenablement les coefficients d'ondelettes. Ce travail est brièvement décrit dans la partie 3.5.3.

2. mais non pertinent !

1.4 Algorithmes d'optimisation non réversible

La totalité des travaux présentés dans cette partie découlent d'un aménagement de l'algorithme de gradient stochastique par Kim-Anh Lê Cao lors des études numériques effectuées dans [2] et [3].

Originellement, l'algorithme de gradient stochastique itératif s'écrit sous la forme :

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

où X_k désigne la position de l'algorithme à l'étape k, γ_k est le pas de l'algorithme, d_k sa direction de descente (aléatoire). Ces algorithmes apparaissent fréquemment en contrôle, théorie du signal ou des images, théorie des jeux, estimation Bayésienne ... Sous des conditions techniques sur la direction d_k et de décroissance du pas de l'algorithme, on peut alors établir les propriétés suivantes.

- Si ζ_k est nul, on est alors ramené par des techniques classiques de martingales (on peut consulter par exemple [Duflo, 1997, Kushner and Yin, 2003][3]) à l'étude de l'équation différentielle ordinaire :
$$dX_t = -\nabla U(X_t)dt.$$
- Lorsque ζ_k est une perturbation gaussienne centrée réduite, le précédent algorithme est plutôt l'approximation « diffusion » (présence du terme $\sqrt{\gamma_k}\zeta_k$) d'une descente de gradient stochastique dès lors que $\mathbb{E}[d_k|\mathcal{F}_k] = -\nabla U(X_k)$ décrite par l'équation différentielle stochastique :
$$dX_t = -\nabla U(X_t)dt + dB_t.$$
On consultera par exemple les ouvrages précédemment cités ainsi que [Benveniste et al., 1990] ou bien [Benaim, 1996] pour une étude plus dynamique de l'algorithme stochastique.

L'aménagement numérique de [2] a alors consisté à simuler une dynamique non markovienne

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

Un tel schéma numérique est fortement relié aux équations différentielles :

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

sous réserve que certaines conditions techniques soient satisfaites et dans le cas où les vecteurs d_k sont en moyenne des directions de descente D. Toutes les études suivantes ont été développées pour étudier de nouvelles méthodes d'optimisation, qu'elles soient déterministes ou bien stochastiques.

1.4.1 Équation différentielle de gradient à mémoire

État de l'art Le premier travail sur ce thème a donc consisté à considérer la famille d'équations différentielles non linéaires qui seraient les limites de l'algorithme (1.8). L'équation d'intérêt a donc été décrite en toute généralité par la « descente à mémoire » :

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

3. Bibliographie réellement non exhaustive.

pour U un potentiel coercif défini sur $\mathbb{R}^d$. En écrivant l'équation à l'ordre 2, on peut alors se ramener par changement de temps (détaillé dans [Cabot, 2009]) à l'étude du système :

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

avec $y = x \circ \tau$, où τ est solution de $\dot{\tau}^2 = k(\tau)/h(\tau)$ et $a = \frac{\dot{k}h + k\dot{h}}{2k^{1/2}h^{3/2}} \circ \tau$. Écrite à l'ordre 2 comme (1.10), l'équation différentielle généralise plusieurs formes connues d'équations. En premier lieu, elle généralise l'équation de Bessel dans le cas où $a(t) = 1/t$ et $U(x) = x^2$ dont les solutions sont les multiples de J_0 donnée par $y(t) \sim \sqrt{\frac{2}{\pi t}}\cos(t - \pi/4)$. On obtient ainsi en temps long $x(t) \sim Ct^{-1/4}\cos(2\sqrt{t} - \pi/4)$.

Dans la communauté d'optimisation convexe, des cas particuliers de telles équations avaient déjà été étudiées dans le cas où a est une fonction constante positive. On retrouve ainsi le système HBF (*Heavy Ball with Friction*) introduit dans [Polyak, 1987] et [Antipin, 1994] pour étudier les propriétés optimisantes des trajectoires. Son étude a alors été généralisé dans un cadre de systèmes dissipatifs par [Hale, 1988, Haraux, 1991] et on peut démontrer la convergence de tels systèmes avec amortissement constant vers des points critiques de U sous des conditions techniques de type convexité à l'infini ou analycité. Enfin, on peut noter les très récents développements de [Ben Hassen and Haraux, 2011] qui propose l'utilisation de l'inertie liée à $\ddot{y}$ avec un amortissement de la forme $g(\ddot{y}(t))$ et de [Haraux, 2007] qui étudie le cas d'un second membre non nul pour une équation similaire à (1.9). L'idée d'adapter le coefficient d'amortissement à la position de la vitesse de la particule peut en effet avoir un intérêt d'un point de vue de l'optimisation mais change alors radicalement l'étude que nous avons effectuée avec un amortissement $a(t)$ dépendant uniquement du temps.

Contributions Le comportement en temps long des trajectoires vérifiant (1.9) ou (1.10) ainsi que le comportement de $U(x(t))_{t\geqslant 0}$ est précisément décrit dans [13] pour des situations où le potentiel est convexe ou du moins sans point d'accumulation dans les parties où $\nabla U = 0$. Par ailleurs, nous donnons également des résultats de convergence dans la situation très particulière uni-dimensionnelle qui ne sont pas généralisables facilement en dimension supérieure. Enfin, des résultats de non convergence sont établis dans [14] pour des situations un petit peu pathologiques où le potentiel possède des parties plates. Des détails sont fournis sur le comportement en temps long de cette équation différentielle « non ordinaire » dans la partie 4.1.

1.4.2 Diffusions à mémoire

État de l'art en processus renforcés L'étude du processus défini par (1.8) lorsque ζ_k est gaussien est numériquement une approximation du régime de diffusion puisque le processus est bruité par un incrément brownien $\sqrt{\gamma_k}d\zeta_k$. Ainsi, il était naturel de s'intéresser à la diffusion basée sur la dérive donnée par (1.9). Le processus diffusif de gradient à mémoire est alors

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

La difficulté probabiliste liée à (1.11) provient principalement de son caractère non Markovien puisque le processus est en interaction avec tout son passé par le biais de la moyennisation en temps de $\nabla U(x_s), 0 \leqslant s \leqslant t$. En un certain sens, ces processus appartiennent donc à la grande classe de processus inter-agissants dont un premier exemple fut introduit par [Coppersmith and Diaconis, 1987] pour les marches aléatoires puis étudié par [Pemantle, 1992] pour étudier un modèle de croissance de polymères, et par [Cranston and Le Jan, 1995] pour les processus en temps continus.

En temps continus, les processus de renforcements sont en général issus d'une convolution entre une fonctionnelle et la mesure d'occupation normalisée dans les travaux de [Benaïm et al., 2002] ou non normalisée comme dans [Durrett and Rogers, 1992]. Il s'agit donc en général d'une interaction à l'instant t mettant en jeu $(X_t - X_s)_{0 \leqslant s \leqslant t}$. Remarquons que le travail [Benaïm et al., 2002] est abordé en utilisant un outil puissant de description de systèmes dynamiques qui sont les pseudo-trajectoires asymptotiques introduits dans [Benaïm and Hirsh, 1996] et qu'une étude aurait pu être mise en oeuvre pour notre système en utilisant ces outils même si le cadre étudiés dans [Benaïm et al., 2002] est compact en espace, ce qui simplifie un certain nombre de difficultés techniques supplémentaires. Enfin, signalons les avancées récentes de [Kurtzman, 2009] qui donne des résultats pour des situations non compactes en adjoignant un potentiel de confinement dans les processus introduits par [Benaïm et al., 2002].

Lien avec certains processus hypoelliptiques Dans [16], nous proposons une technique de grossissement d'espace pour rendre la diffusion Markovienne au prix d'une présence de dégénérescence forte dans le système aléatoire sur la seconde coordonnée. En définissant $(Y_t)_{t \geqslant 0}$ comme le processus donné par le drift de (1.11) à l'instant t et en posant $r = \dot{k}/k$, on ramène alors l'étude de la diffusion à

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

Ces équations couplées (1.12) tombent alors dans le registre des travaux sur les processus hypoelliptiques. De très nombreuses avancées ont été faites sur ce sujet ces dernières années, notamment à la suite de travaux comme ceux de [Helffer and Nier, 2005] et [Villani, 2009] qui s'intéressent au comportement en temps long de processus comparables.

La difficulté principale pour l'étude des convergences à l'équilibre de tels systèmes réside en le fait que ces convergences ne peuvent pas être déduites directement d'inégalités fonctionnelles classiques associées au critère Γ_2 de [Bakry and Émery, 1985] par exemple. Un cas particulier de telles équations, les équations de Fokker-Planck cinétiques, a peut être reçu un peu plus d'intérêt que d'autres comme l'atteste les nombreuses références à son sujet. On pourra consulter par exemple les travaux de [Risken, 1989, Eckmann and Hairer, 2003, Hérau and Nier, 2004] qui abordent ces équations en étudiant attentivement le spectre de l'opérateur tandis que d'autres travaux [Desvillettes and Villani, 2001, Dolbeault et al., 2009] utilisent des constructions de normes un peu plus coercives (qui sont des fonctions de Lyapunov du sytème dynamique) que la norme $\mathbb{L}^2$ pour obtenir l'applicabilité du Lemme de Gronwall. On peut également remarquer que le lien entre inégalités fonctionnelles et fonctions de Lyapunov a été démontré dans [Bakry et al., 2008] mais que dans le contexte hypo-elliptique, un tel lien ne suffit pas toujours pour obtenir des vitesses de convergence à l'équilibre.

Enfin, il faut noter que le contexte hypo-elliptique induit également des difficultés techniques concernant l'existence et la régularité des lois $P_t(z_0, .)$ où t est le temps courant du processus et z_0 son point d'initialisation au temps $t = 0$. C'est dans le contexte des travaux de Hormander et son théorème sur les sommes de carrés de champs de vecteurs que ces résultats sont à rechercher. De très nombreuses avancées ont eu lieu dans ce domaine depuis les travaux initiaux de [Hörmander, 1967], certains font intervenir des outils provenant des équations aux dérivées partielles comme [Kohn, 1978, Trèves, 1980], d'autres du calcul de Malliavin (voir [Kusuoka and Stroock, 1987, Cattiaux, 1992, Hairer, 2011]).

Par ailleurs, sous des conditions de contrôle des trajectoires, il est même possible de trouver des estimations précises des densités par le biais de calcul de Malliavin [Delarue and Menozzi, 2010, Bally and Kohatsu-Higa, 2010] ou d'inégalités de type Harnack [Pascucci and Polidoro, 2006,

Polidoro, 1997]. Bien entendu ces conditions de contrôle (plus de précisions seront fournies plus loin) sont loin d'être innocentes puisqu'elles sont déjà nécessaires pour assurer des propriétés de positivité du semi-groupe par le biais de l'application du théorème du support (voir [Stroock and Varadhan, 1972]). On pourra également consulter [Ben Arous and Léandre, 1991] qui fournissent une condition nécessaire et suffisante sous certaines hypothèses de bornitudes des coefficients de champs de vecteur pour assurer une telle stricte positivité.

Contributions Nous étudions dans [16] les propriétés de stabilité des processus de gradients moyennés définis par les équations couplées (1.12).

Sous des conditions techniques sur U (principalement U doit être « convexe » à l'infini avec $U(x)/|x| \to +\infty$), nous démontrons que le système élargi (1.12) est entièrement caractérisé par le comportement en temps long de $r(t) = \dot{k}(t)/k(t)$. En particulier, on démontre la stabilité d'un tel processus lorsque la mémoire du processus est suffisament courte, tandis que le processus à une tendance naturelle à exploser dès que la mémoire s'effectue à plus long terme.

Les difficultés principales concernent tout d'abord la preuve de l'hypo-ellipticité du processus ainsi que sa controllabilité approchée. Par ailleurs, la stabilité de ce processus passe par la construction d'une fonction de Lyapunov non triviale permettant à la fois de controler le processus en vitesse et position. Enfin, il est possible d'obtenir des vitesses de convergence en variation totale de la mesure d'occupation du processus vers sa mesure stationnaire en utilisant conjointement des arguments de type Lyapunov avec des arguments de régularité en s'inspirant des travaux de [Down et al., 1995]. Ces vitesses sont rendues relativement explicites dans notre cas en utilisant alors les avancées récentes de [Douc et al., 2009]. Je détaille dans le paragraphe 4.2 l'étude menée sur le système (1.11)-(1.12).

1.4.3 Lien avec les équations de Fokker-Planck cinétiques

L'écriture du système (1.12) sous forme couplée est fortement évocatrice des équations de Fokker-Planck cinétiques décrites par

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

Il y a néanmoins une différence substantielle qui est que la mesure stationnaire du processus (1.13) est connue alors que celle du gradient moyennée ne l'est pas, mis à part quelques situations très particulières comme lorsque U est un potentiel quadratique ($U(x) = a|x|^2$). Dans ce cas précis, le processus défini par (1.12) est gaussien et il est alors facile d'identifier sa mesure stationnaire. Néanmoins, les deux systèmes (1.12) et (1.13) ne semblent pas équivalents (ou du moins aucune équivalence immédiate ne peut être déduite pour l'instant) en toute généralité.

Le processus déterministe donné par (1.9) possède des propriétés optimisantes intéressantes, et c'est ce qui nous a conduit à étudier sa version perturbée par un bruit gaussien à petit paramètre. Plus que la stabilité du processus de gradient moyenné subsiste réellement la nature de la vitesse de convergence qui peut renseigner sur la faculté qu'aura le processus (1.12) à également se comporter efficacement pour des méthodes d'optimisation.

Contributions L'étude [16] n'ayant apporté que des réponses partielles en terme de vitesse de convergence (uniquement des vitesses en variation totale), nous abordons dans [15] l'étude précise du spectre de l'opérateur de Fokker-Planck cinétique (1.13) puisque les calculs semblent plus abordables que ceux de (1.12). Nous calculons alors les vitesses de convergence en norme L^2

pour des cas particuliers de potentiels. Notamment, nous donnons des vitesses (et équivalents asymptotiques en temps long et en temps petit) optimales dans des cas très simples où $U = ax^2/2$, et $U = 0$ sur le tore $\mathbb{T} = [0;1]$ pour le processus (1.13). Ces estimations sont basées sur des méthodes différentes de celles fournies dans [Dolbeault et al., 2009] ou [Villani, 2009] qui utilisent une autre décomposition de l'opérateur de Fokker-Planck cinétique que celle décrite dans [15]. Les résultats obtenus ainsi que les outils mis en oeuvre pour ces objectifs sont détaillés dans le paragraphe 4.3.

1.4.4 Diffusion moyennée à petit paramètre

La détermination fine des vitesses de convergence en temps long décrite dans le paragraphe précédent n'a rien d'innocente puisque l'objectif avoué est à terme de mettre en oeuvre un algorithme d'optimisation stochastique du potentiel U. Celui-ci pourrait alors découler d'un recuit simulé utilisant soit (1.12), soit (1.13) en prenant alors un processus dont le paramètre de diffusion $\sigma(t) \longmapsto 0$ lorsque $t \longmapsto +\infty$. Nous noterons par la suite $\sigma = \epsilon$ pour faire référence à des diffusions à petit paramètre.

Algorithme de recuit simulé Dans le cas standard de la diffusion elliptique à valeurs dans $\mathbb{R}^n$

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

il est bien connu (on se référera par exemple à [Miclo, 1992]) que le processus peut permettre de minimiser *globalement* le potentiel U dès lors que $\epsilon(t) \longmapsto 0$ à une vitesse appropriée. L'efficacité de tels algorithmes de recuit simulé dépend à la fois de

1. la vitesse de convergence du processus vers sa mesure d'équilibre μ_ϵ lorsque ϵ est constant
2. la vitesse de convergence de μ_ϵ vers μ_∞ lorsque $\epsilon \mapsto 0$.

Notamment, cet équilibre entre les deux vitesses permet de trouver le schéma de température le plus approprié pour effectuer le recuit simulé (plus la vitesse de convergence du processus à ϵ constant vers sa mesure d'équilibre est rapide, plus on peut accélérer le schéma de décroissance de température $\epsilon(t) \longmapsto 0$) et meilleur est l'algorithme.

Plus précisément, lorsque ϵ est constant et dans les situations de convergence à l'équilibre μ_ϵ, on peut espérer dans le cas diffusif « pur » des vitesses $\mathbb{L}^2$ exponentielles

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

où la constante $A(\epsilon)$ joue alors un rôle majeur dans la calibration de la température de $t \mapsto \epsilon(t)$. En effet, [Miclo, 1992, Chiang et al., 1987, Royer, 1989] montrent alors qu'il existe $d^* > 0$ optimale tel que $\epsilon(t) = c/\ln(t)$ assure la convergence du processus dès que $c > d^*$ vers $\min_{\mathbb{R}^n} U$. Cette constante d^* n'est pas connue par le praticien puisqu'elle dépend d'une connaissance du potentiel U inaccessible en pratique. Ainsi, il s'agit de trouver une manière d'évaluer une majoration de d^* qui permet alors, de choisir un schéma de température admissible, et l'obtention d'une fonction $A(\epsilon)$ assez grande est importante pour la mise en place du recuit simulé.

Si on considère l'approche de [Bakry et al., 2008], dès lors qu'il existe une fonction de Lyapunov appropriée, il y a alors une inégalité de Poincaré avec une constante C_P qui n'est pas forcément optimale [4]. Par ailleurs, les vitesses de convergence sont alors de la forme

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

4. en général la constante C_P qu'on exhibe par le biais des fonctions de Lyapunov est trop grande

L'utilisation d'une telle approche permet alors de donner une constante admissible pour le recuit simulé mais qui est trop petite pour être optimale puisque la constante C_P trouvée est trop grande. Il est également possible d'étudier directement le comportement du spectre de l'opérateur Markovien L_ϵ issu de (1.14) pour les petites valeurs de ϵ (voir par exemple la section 7 du chapitre 6 de [Freidlin and Wentzell, 1984] pour le cas des variétés compactes contenant un point stable). Bien entendu, lorsque $\epsilon \longmapsto 0$, la plus petite valeur propre de $-L_\epsilon$ se comporte en $\exp(-\Delta V/\epsilon^2)$ où ΔV est une constante explicite dépendant du quasi-potentiel associé à la fonctionnelle de grande déviation de (1.14). À nouveau, la connaissance de cette constante est inaccessible en pratique lors d'une expérimentation pour laquelle les connaissances sur U sont partielles, ainsi la calibration pratique de ϵ ne peut être déduite de ce genre d'approche.

Modèles du second ordre Plutôt qu'essayer d'estimer assez vainement une constante $A(\epsilon)$ (ou $d^\star$) dans (1.15), il est imaginable en pratique de sortir du carcan de la diffusion standard (1.14) et construire une autre diffusion telle que la convergence à l'équilibre se produit, et qui possède une contractivité $\tilde{A}(\epsilon)$ naturellement plus grande.

Il est tentant d'essayer utiliser des modèles du second ordre puisque ceux-ci peuvent avoir une capacité exploratoire plus importante que certains du premiers ordre (phénomène déjà observé par exemple dans le cas déterministe pour le système dynamique (1.9)). Par ailleurs, [Diaconis et al., 2010b] observe un tel phénomène dans une situation discrète de chaîne de Markov d'ordre 2. Dans [21], nous étudions donc le comportement des diffusions moyennées à petits paramètres (sans pour autant aller jusqu'à la description d'un algorithme de recuit simulé basé sur (1.9)). La première question a été d'identifier s'il y avait un comportement limpide de la mesure stationnaire ν_ϵ de (1.11) lorsque ϵ devient petit. En effet, c'est avant tout le comportement de ν_ϵ pour ϵ petit qui peut assurer ou non qu'un processus (1.11) se concentre sur les minima globaux de U.[5]

Dans le travail [21], nous étudions donc le cas de l'équation de diffusion moyennée (1.11) homogène en temps correspondant à des mémoires $h(t) = k(t) = e^{\lambda t}$ et démontrons un principe de Grandes Déviations en $\epsilon \to 0$ pour ν_ϵ la mesure invariante du processus (qui est cette fois markovien homogène du fait de ce choix particulier de fonctions mémoires). Mis à part le cas spécifique du potentiel $U(x) = ax^2/x$, il n'existe pas de formule explicite de la mesure invariante ν_ϵ. Par conséquent, même la nature du quasi-potentiel dérivant du principe de Grandes Déviations précédent est ambigüe. Nous donnons dans [21] des conditions suffisantes sur U (et sans doute non optimales) pour que ν_ϵ se concentre sur le minimum global de U. Les détails de ces travaux sont décrits dans le paragraphe 4.4.

5. En ce sens, une étude sur les équations de Fokker Planck cinétiques aurait été « plus simple » (même si le processus est markovien non réversible également) principalement parce que la mesure invariante associée à (1.13) est explicite $m_\epsilon(x,v) \propto e^{-[U(x)+v^2/2]/\epsilon^2}$ et que le comportement en $\epsilon \longmapsto 0$ de sa marginale en x est évident par le biais de la méthode de Laplace.

Chapitre 2

Modélisation et statistiques en grande dimension

Dans le contexte de ce chapitre, nous étudions des problèmes où on dispose des observations $(X_i, Y_i)_{i=1...n}$ tels que les X_i sont décrits par p variables d'un dictionnaire $\mathcal{D} = (g_1, \ldots, g_p)$. Les Y_i sont des labels qui codent la classe d'appartenance de la donnée X_i pour des problèmes de classification, ou tout simplement les Y_i sont des éléments de $\mathbb{R}^d$ lorsqu'on étudie des problèmes de régression. Notre étude se placera toujours dans le contexte où $p >> n$, ce qui rend difficile les approches standard en raison du piège de la grande dimension.

2.1 Algorithme stochastique de sélection de variables

Dans le contexte de la classification supervisée, il est usuel d'avoir à sa disposition un algorithme de classification (noté génériquement $\mathbb{A}$ dans cette section) et notre objectif est ici de type « best subset », à savoir trouver $\mathcal{G} \subset \mathcal{D}$ tel que la prédiction de $\mathbb{A}$ en utilisant les variables de $\mathcal{G}$ est optimale.

2.1.1 Description du modèle

Pour la suite, on définit $\hat{\mathbb{A}}_{\mathcal{G},n}$ l'apprentissage de l'algorithme $\mathbb{A}$ en utilisant les données $(X_i, Y_i)_{i=1...n}$ et les variables actives de $\mathcal{G}$. Si on définit l'erreur de prédiction de $\hat{\mathbb{A}}_{\mathcal{G},n}$ par

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

l'approche idéale consisterait à trouver

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

Bien entendu, résoudre (2.1) est impossible numériquement puisque il s'agit d'énumérer tous les sous-ensembles de $\mathcal{D}$ qui est un problème NP difficile. Par ailleurs, la loi jointe (X, Y) étant inconnue, il n'est pas possible en réalité de mesurer q exactement mais seulement en donner une estimation sur l'ensemble d'apprentissage qu'on notera $\hat{q}$ (voir [6] pour plus de détails sur la stratégie boostrap mise en place).

L'approche développée a été de proposer une stratégie (certainement sous-optimale mais calculable numériquement) de pondération des éléments de $\mathcal{D}$ par une loi de probabilité discrète $\mathbb{P}$ qui sera l'inconnue de notre problème. Nous fixons un entier k plus petit que n, et associons

à $\mathbb{P}$ une énergie qui quantifie l'erreur de prédiction en moyenne lorsque les variables de $\mathcal{D}$ sont tirées avec remise sous la loi $\mathbb{P}$:

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

On constate assez facilement que si $\mathbb{P}$ est proche du minimum de $\mathcal{E}$, alors la distribution de poids sur $\mathcal{D}$ va privilégier les variables qui permettent de bien classer les données au travers de $\mathbb{A}$. Il est donc naturel de procéder à un algorithme d'optimisation de $\mathcal{E}$. Notons que $\mathcal{E}$ est une fonction polynomiale de degré k de $\mathbb{P}$ dont les coefficients ne sont pas connus (sinon cela réclamerait de connaître toutes les valeurs de $\hat{q}(\hat{\mathbb{A}}_{\mathcal{G},n})$, ce qui est à nouveau NP difficile).

2.1.2 Algorithme de descente de gradient

Nous proposons dans [6] de minimiser $\mathcal{E}$ au travers d'une stratégie itérative de descente de gradient éventuellement perturbée par une diffusion à petit paramètre. Le schéma global de l'algorithme est décrit dans la figure 2.1.

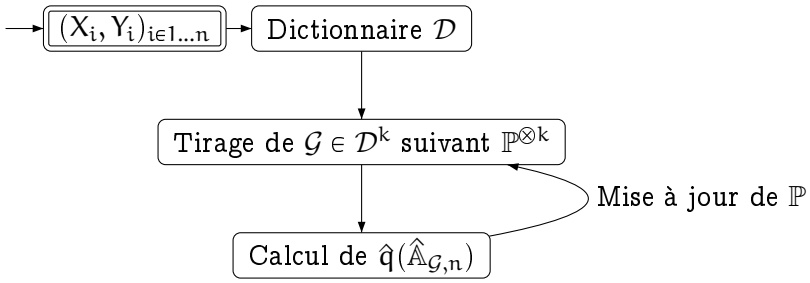


FIGURE 2.1 – Schéma itératif d'apprentissage de $\mathbb{P}$.

Pour tout point $\mathbb{P}$ appartenant au simplexe $\mathcal{S}_\mathcal{D}$ des mesures de probabilités discrètes sur $\mathcal{D}$, on peut calculer le gradient euclien de $\mathcal{E}$:

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

où $C(\mathcal{G}, g)$ désigne simplement le nombre d'occurrences de g dans $\mathcal{G}$. Ainsi, si $\pi_\mathcal{D}$ désigne la projection sur l'hyperplan $\mathcal{H}_\mathcal{D}$ qui contient $\mathcal{S}_\mathcal{D}$, un algorithme d'optimisation de $\mathcal{E}$ pourrait consister à effectuer

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

tandis que la diffusion à petit paramètre associée à cette descente de gradient serait définie au travers de la diffusion sous contrainte

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

Par la suite, on notera $\pi_\mathcal{S}$ la projection sur le simplexe $\mathcal{S}_\mathcal{D}$ puisque celle-ci sera nécessaire pour mettre en oeuvre notre algorithme d'apprentissage.

Dans l'équation (2.5), $(B_t)_{t\geqslant 0}$ est un mouvement brownien sur $\mathbb{R}^p$, $\sigma_\mathcal{D}$ est une matrice de covariance issue de la projection sur $\mathcal{H}_\mathcal{D}$ et dZ_t est un processus de saut introduit pour contraindre $(\mathbb{P}_t)_{t\geqslant 0}$ à être à tout instant une mesure de probabilité sur $\mathcal{D}$. On ne rentrera

pas dans les considérations techniques qui garantissent l'existence et l'unicité de solutions de l'équation (2.5). Ces résultats sont prouvés dans [5] à partir de l'application de Skorokhod décrite dans [Dupuis and Ramanan, 1999].

2.1.3 Approximation par gradient stochastique

Notre idée est de rendre chaque itération de l'algorithme dépendante d'un seul calcul $\hat{q}(\hat{\mathbb{A}}_{\mathcal{G},n})$, ce qui n'est pas trivial puisque l'expression (2.3) montre qu'un calcul exact de $\nabla\mathcal{E}(\mathbb{P})$ n'est possible que si on énumère à nouveau tous les k-uplets de variables. En réalité, si on fait bien attention à la nature de $\mathcal{E}$ et de son gradient, il est possible d'écrire que

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

Il est alors possible de produire deux algorithmes stochastiques approchant les comportements de (2.4) et (2.5), ce qui rend possible l'apprentissage de $\mathbb{P}$. Étant donnée une suite de pas décroissants $(\alpha_j)_{j\in\mathbb{N}}$ satisfaisant les deux conditions techniques

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

on définit alors la méthode d'apprentissage de la suite de probabilités $(\mathbb{P}_j)_{j\geqslant 0}$ (décrit par l'algorithme (1)).

Algorithm 1 Sélection de variable par gradient stochastique (approximation de (2.4)).

Require: Dictionnaire $\mathcal{D}$, Algorithme $\mathbb{A}$, Données $(X_i, Y_i)_{i\in 1\ldots n}$, entiers $k\in]0;n[$ et J.
Ensure: $\mathbb{P}$ minimiseur de $\mathcal{E}$
  $\mathbb{P}_0 = \mathcal{U}_{\mathcal{D}}$, loi uniforme sur $\mathcal{D}$.
  $j \leftarrow 0$
  while $j < J$ **do**
    Tirer $\mathcal{G}_j$ dans $\mathcal{D}^k$ selon $\mathbb{P}_j^{\otimes k}$
    Calculer $\hat{\mathbb{A}}_{\mathcal{G}_j,n}$ ainsi que l'estimation de l'erreur de classification $\hat{q}(\hat{\mathbb{A}}_{\mathcal{G}_j,n})$
    Mise à jour de la distribution de poids

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

    $j \leftarrow j+1$
  end while

En écrivant le processus affine par morceaux $(\mathbb{P}_t^{\text{interp}})_{t\geqslant 0}$ interpolant $(\mathbb{P}_j)_{j\geqslant 0}$ aux temps

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

on montre alors en s'inspirant des résultats de gradient stochastique de Robbins-Monro (voir par exemple [Kushner and Yin, 2003] ou [Benaim, 1996]) le résultat suivant :

Théorème 2.1.1 (Convergence de l'algorithme OFW (*Optimal Feature Weighting*)) *Le processus stochastique $(\mathbb{P}_t^{\text{interp}})_{t\geqslant 0}$ est une pseudo-trajectoire asymptotique du flot de l'équation différentielle (2.4). Par ailleurs, l'algorithme converge vers un minimum (local) de $\mathcal{E}$.*

Il est également possible de mener une étude sur la diffusion pour l'approximation stochastique précédente (c'est une partie des résultats présentés dans [5]). Cette seconde méthode plus exploratoire est décrit par l'algorithme (2).

Algorithm 2 Sélection de variable par diffusion stochastique (approximation de (2.5)).

Require: Dictionnaire $\mathcal{D}$, Algorithme $\mathbb{A}$, Données $(X_i, Y_i)_{i\in 1\dots n}$, entiers $k \in]0; n[$ et J, variance σ^2.

Ensure: $\check{\mathbb{P}}$ minimiseur de $\mathcal{E}$

$\check{\mathbb{P}}_0 = \mathcal{U}_{\mathcal{D}}$, loi uniforme sur $\mathcal{D}$.

$j \leftarrow 0$

while $j < J$ **do**

Tirer $\mathcal{G}_j$ dans $\mathcal{D}^k$ selon $\check{\mathbb{P}}_j^{\otimes k}$

Calculer $\hat{\mathbb{A}}_{\mathcal{G}_j,n}$ ainsi que l'estimation de l'erreur de classification $\hat{q}(\hat{\mathbb{A}}_{\mathcal{G}_j,n})$

Tirage de p variables aléatoires indépendantes $(\xi_j(g))_{g\in\mathcal{D}} \sim \mathcal{N}(0,1)^{\otimes p}$

Mise à jour de la distribution de poids

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

$j \leftarrow j + 1$

end while

À nouveau, en définissant $(\check{\mathbb{P}}_t^{\text{interp}})_{t\geqslant 0}$ le processus affine par morceaux interpolant $(\check{\mathbb{P}}_j)_{j\geqslant 0}$ aux temps τ_j, et en utilisant les méthodes classiques d'approximation stochastique de [Kushner and Yin, 2003] et [A. Benveniste and Priouret, 1987] de tension/identification, on peut démontrer le résultat suivant.

Théorème 2.1.2 (Convergence de l'algorithme de diffusion OFW) *Le processus stochastique $(\check{\mathbb{P}}_t^{\text{interp}})_{t\geqslant 0}$ converge faiblement vers l'unique mesure invariante du processus markovien (2.5). Il en est de même $(\check{\mathbb{P}}_j)_{j\in\mathbb{N}}$.*

À noter que dans ce résultat, la difficulté consiste à démontrer la tension du processus lorsque celui-ci est effectivement réfléchi sur le simplexe $\mathcal{S}_{\mathcal{D}}$.

La figure 2.2 représente à titre d'exemple les détecteurs binaires de bords sélectionnés lors d'un problème de détection de visages avec un algorithme de SVM.

2.2 Algorithme séquentiel de plans d'expériences

Nous proposons dans cette partie un nouvel algorithme séquentiel pour trouver des plans d'expériences adptatifs dans le contexte de la régression. Pour des raisons de simplicité, on se cantonnera au cas où la fonction η inconnue est définie sur $\Omega = [0; 1]^d$ et on cherche un algorithme permettant de choisir des points de mesure en nombre fini pour prédire ensuite au mieux la fonction η sur Ω. Dans la suite, nous décrivons l'approche $d = 1$ et ceci sans perte de généralités pour l'algorithme et les résultats théoriques qui lui sont associés.

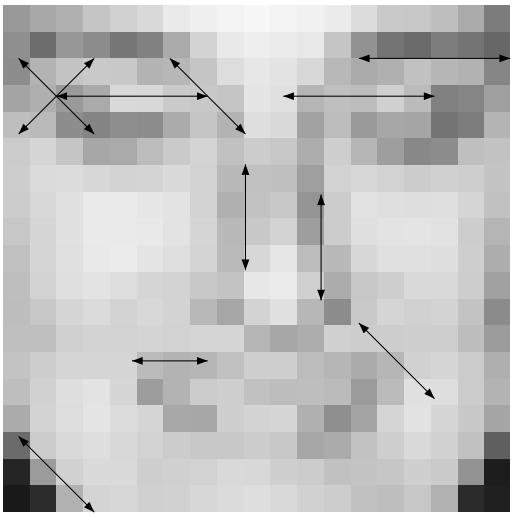


FIGURE 2.2 – Représentation des détecteurs de bords binaires sélectionnés pour un problème de classification de visage.

2.2.1 Cadre

La fonction η est supposée appartenir à un espace de Besov homogène de régularité s inconnue [1]. Nous cherchons à prédire η comme combinaison linéaire finie d'éléments $(\Lambda_{j,k})_{j\in\mathbb{N},k=0\ldots2^j-1} = \mathcal{D}$. Ici, $\mathcal{D}$ est donc une analyse multi-résolution basée sur une décomposition en ondelettes et nous supposons que le signal η est mesuré par l'intermédiaire d'observations f telles que

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

où $\xi(x)$ est un bruit gaussien centré réduit et σ^2 la variance du bruit a priori inconnue. Nous cherchons donc à choisir les points de mesure utilisant la formule (2.6) pour lesquels la prédiction de η sera optimale.

Dans cet algorithme, nous proposons d'utiliser un modèle linéaire basé sur un sous-ensemble d'éléments de $\mathcal{D}$. La difficulté revient alors à fixer à chaque étape n le plan d'expérience $(\mathbf{x}_n)$ ainsi que la base utilisée $\mathcal{D}_n$. Pour des raisons évidentes liées à la problématique des plans d'expériences [2], on impose donc

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

Nous associons à notre estimation par modèle linéaire $\hat{\eta}_{\mathbf{x}_n,\mathcal{D}_n}$ un critère qualitatif d'interpolation naturel qui correspond à l'erreur quadratique moyenne intégrée

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

η se décompose alors sur l'espace engendré par $\mathcal{D}_n$ et son orthogonal et le critère s'écrit classiquement en « biais + variance » :

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

Même si le biais est inacessible, il est possible d'en donner une valeur pessimiste par une approche minimax dépendant d'un paramètre $\tau > 0$ quantifiant l'amplitude du biais incrompressible dû

1. Ici, le caractère adaptatif de l'algorithme n'a rien de commun avec l'estimation adaptative que l'on peut rencontrer dans certains travaux de statistiques mathématiques

2. chaque mesure de η étant considérée comme couteuse, on ne « jette » pas une mesure une fois que son prix a été payé

à l'utilisation d'éléments de $\mathcal{D}_n$:

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

Ces simples considérations conduisent alors à considérer le critère minimax équilibré :

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

où $\lambda = \sigma^2\tau^{-2}$ est en fin de compte un paramètre de pénalisation de la variance.

2.2.2 Description de l'algorithme de plans d'expériences séquentiels

Nous proposons dans [4] un algorithme itératif construisant le plan $\mathbf{x}_n$ et proposant ensuite une mise à jour de $\mathcal{D}_n$: $\mathbf{x}_n$ permet de contrôler la variance tandis que $\mathcal{D}_n$ optimise le terme de biais. La détermination de ζ_{n+1} (voir (2.7) se fait par l'optimisation de $J^\star(\hat{\eta}_{\mathbf{x}_n,\mathcal{D}_n}, \eta)$ tandis que $\mathcal{D}_{n+1}$ s'obtient en ajoutant ou soustrayant à $\mathcal{D}_n$ un de ses fils par le biais d'une stratégie Metropolis-Hastings de sorte que $|\mathcal{D}_{n+1} \Delta \mathcal{D}_n| = 1$. Cette méthode est décrit dans l'algorithme 3.

Algorithm 3 Construction séquentielle de plan d'expérience.

Require: Dictionnaire $\mathcal{D}_0$, paramètre $\lambda \in [0; +\infty]$, budget de mesure n.
Ensure: $\mathbf{x}_n$ et $\mathcal{D}_n$
  Détermination de $\mathbf{x}_0$ plan optimal minimisant (2.8).
  Mesurer f par le biais de (2.6) et calcul du modèle linéaire $\hat{\eta}_{\mathbf{x}_0,\mathcal{D}_0}$.
  $j \leftarrow 0$
  while $j < n$ **do**
    Mise à jour de $\mathcal{D}_{j+1}$ en choisissant aléatoirement
  – Ajoût d'un fils ou parent d'un élément de $\mathcal{D}_j$ le plus significatif
  – Suppression d'un fils ou d'un parent le moins significatif de $\mathcal{D}_j$
  – Ne pas modifier $\mathcal{D}_j$
    Calculer ζ_{j+1} par minimisation de (2.8).
    Mesurer $f(\zeta_{j+1})$ par le biais de (2.6) et mettre à jour le modèle linéaire $\hat{\eta}_{\mathbf{x}_{j+1},\mathcal{D}_{j+1}}$.
    $j \leftarrow j + 1$
  end while

L'idée est donc de coupler à une méthode forward/backward stochastique de sélection de variables une stratégie de plan d'expérience. La description précise du choix de $\mathcal{D}_{j+1}$ en fonction de $\mathcal{D}_j$ est un petit peu fastidieuse et je renvoie à [4] pour plus de détails sur cette mise à jour. Bien entendu, afin de rendre adaptatif l'algorithme aux mesures déjà effectuées, cette mise à jour de $\mathcal{D}_j$ dépend du modèle linéaire courant $\hat{\eta}_{\mathbf{x}_{j+1},\mathcal{D}_{j+1}}$.

2.2.3 Résultats

Dans l'algorithme précédent, la phase de recherche de ζ_{j+1} est le coeur de la difficulté numérique pour la problématique du plan d'expérience adaptatif puisqu'en général il n'existe pas de formule explicite pour le minorant de (2.8). Dans [4], nous démontrons un résultat de localisation quasi-explicite pour la détermination de ζ_{j+1} pour une situation restrictive où seule la variance du plan est prise en compte ($\lambda = +\infty$) et pour la base des triangles de Schauder. Ce résultat est donné par le théorème suivant.

Théorème 2.2.1 *Quel que soit $\tilde{\mathcal{D}}$ un sous-ensemble fini de $\mathcal{D} = (\Lambda_{j,k})_{j=0\cdots+\infty,k=0\ldots2^j-1}$ ainsi qu'un plan d'expérience préliminaire $\boldsymbol{x}$, le plan optimal $\boldsymbol{x} \cup \zeta$ pour le critère*

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

est obtenu lorsque ζ est localisé sur les points critiques des éléments de $\tilde{\mathcal{D}}$, c'est à dire

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

Ce théorème précédent revêt une importance capitale pour l'algorithme précédent puisqu'il permet de trouver le plan $\mathbf{x}_{j+1}$ en au plus $|\mathcal{D}_{j+1}|$ itérations.

Par ailleurs, dans la situation où $\mathcal{D}_j$ reste fixe tout au long de l'algorithme, il est possible de prouver un résultat théorique sur les coefficients du modèle linéaire pour n'importe quelle base multi-résolution.

Théorème 2.2.2 *Quel que soit $\tilde{\mathcal{D}}$ un sous-ensemble fini de $\mathcal{D} = (\Lambda_{j,k})_{j=0\cdots+\infty,k=0\ldots2^j-1}$, et si $\eta = \eta_{\tilde{\mathcal{D}}} + (\eta - \eta_{\tilde{\mathcal{D}}}$ désigne la décomposition de η sur les éléments de $\tilde{\mathcal{D}}$, alors dans le cas où $\lambda = +\infty$, il existe $C > 0$ tel que*

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

Même si les résultats théoriques ne concernent qu'une base particulière d'ondelettes, il est possible d'utiliser d'autres bases comme celle de Meyer. Sur un exemple particulier, l'efficacité de la méthode de régression adaptative est assez bluffante comme le montre la figure (2.3) puisque seulement une dizaine de points suffisent à capter l'essentiel de l'information pour estimer η. À noter que nous donnons dans [4] des comparaisons expérimentales avec des méthodes plus sophistiquées de seuillages de coefficients d'ondelettes ou de modèles linéaires pénalisés (Lasso) qui démontrent que notre méthode se compare assez favorablement à ces techniques.

2.3 Boosting multivarié : application à la reconstruction de réseaux de régulation

[3]

Comme indiqué dans le paragraphe introductif, le problème d'estimation du réseau de régulation peut être modélisé par un cadre de régression multivariée. Étant donné un espace de Hilbert H, on cherche à approcher $f = (f^1, \ldots f^m) \in H^{\otimes m} := H_m$ par une suite d'éléments $(G_k)_{k\geqslant 0}$. Pour ce faire, on dispose d'un dictionnaire fini de taille p noté $\mathcal{D}$ d'éléments de H tel que $\overline{\mathrm{Span}\,\mathcal{D}} = H$. À la vue du contexte applicatif statistique, nous considérons le cas où $\mathcal{D}$ n'est pas un dictionnaire orthogonal d'élements de H.

2.3.1 Description (sommaire) des algorithmes de boosting

Cadre déterministe Les algorithmes de $\mathbb{L}^2$-Boosting unidimensionnels et déterministes fonctionnent ainsi : la suite de prédicteur G_k de f est initialisée à $G_0 = 0$ et G_k s'obtient à partir de G_{k-1} en lui ajoutant une composante sélectionnée à partir d'un critère basé sur $f - G_{k-1}$ et $\mathcal{D}$. Nous précisons dans l'Algorithme 4 le cas particulier du *Weak Greedy Algorithm* même s'il existe un grand nombre de variantes plus ou moins performantes de telles méthodes de Boosting.

3. Ce travail constitue une partie de la thèse de Magali Champion débutée en 2011.

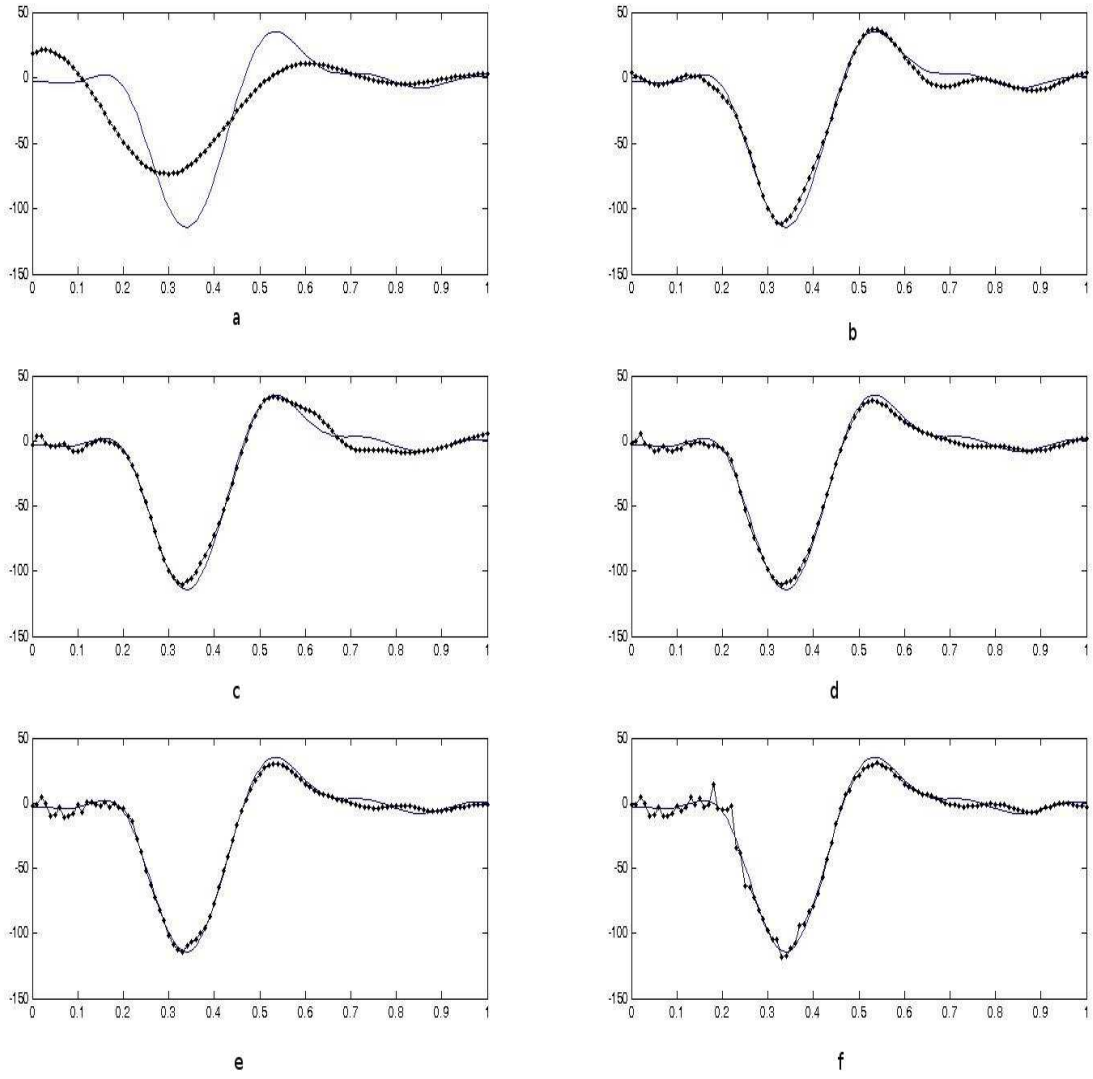


FIGURE 2.3 – Régression obtenue par notre algorithme séquentiel de plan d'expérience sur la base « Motorcycle » avec 5 points (a), 15 points (b), 25 (c), 35 (d), 45 (e), 55 (f). Courbe continue : vrai signal, Courbe pointillée : interpolation avec un modèle linéaire et un sous dictionnaire de la base de Meyer.

Bien entendu, la capacité d'un tel algorithme à approcher f dépend de la « taille »de f. Les travaux [DeVore and Temlyakov, 1996] permettent alors de donner la vitesse de convergence de G_k vers f et la taille de f est définie au travers de la constante B dans le résultat suivant :

Théorème 2.3.1 ([DeVore and Temlyakov, 1996]) *Soit* $B > 0$ *et supposons* $f \in \mathcal{A}(\mathcal{D}, B)$ *où*

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

alors il existe une constante C_B *ne dépendant que de* B *telle que le résidu d'approximation satisfait*

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

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

Require: Dictionnaire $\mathcal{D}$, fonction $f \in H$ à approcher.

Ensure: Paramètre de *shrinkage* $\nu \in]0,1]$, Itération maximale N

Prédicteur $G_0 = 0_H$ et Résidu $R_0 = f$.

$k \leftarrow 0$

while $k < N$ **do**

Choix d'un élément $\varphi_k \in \mathcal{D}$ suffisament corrélé avec R_k $|\langle R_k, \varphi_k \rangle| \geqslant \nu \max_{g \in \mathcal{D}} |\langle R_k, g \rangle|$

Mise à jour de la prédiction

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

et des résidus

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

$k \leftarrow k + 1$

end while

L'effet de ν est de ralentir la vitesse de convergence de l'algorithme qui est optimale pour $\nu = 1$: on obtient alors une vitesse en $k^{-1/6}$. Même si $\nu < 1$ semble ici inutile, il est en réalité très important en vue d'une application de l'algorithme de Boosting dans le cadre aléatoire (voir plus bas).

Cadre aléatoire Les travaux de [Bühlmann, 2006] exploitent l'algorithme précédent dans une situation de régression bruitée et montrent la stabilité de tels algorithmes dans le cas où on dispose d'un échantillon $(X_1, Y_1), \dots (X_n, Y_n)$ i.i.d. où

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

en supposant à nouveau que $f \in \overline{\operatorname{Span} \mathcal{D}}$. En notant H l'espace de Hilbert $\mathbb{L}^2(P)$ où P désigne la loi (inconnue) des X, on est dans une situation ici où l'on ne peut mesurer qu'empiriquement sur les données les tailles de certains éléments de H. On note donc

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

Le *Weak Greedy Algorithm* s'étend au cadre bruité et est décrit par l'algorithme 5 [4].

2.3.2 Algorithme de Boost-Boost pour le cadre multivarié déterministe

Extension de [Lutz and Bühlmann, 2006] Dans la situation où l'élément à prédire f est m dimensionnel, il est envisageable d'étendre les algorithmes précédents en considérant (par exemple dans la situation déterministe) à nouveau une suite de prédicteurs/résidus initialisés en $G_0 = 0_{H_m}, R_0 = f$ et on choisit à l'itération k la coordonnée $i_k \in \{1 \dots m\}$ et le prédicteur $\varphi_k \in \mathcal{D}$ tels que

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

Un tel choix de coordonnée et prédicteur a été proposée par [Lutz and Bühlmann, 2006] dans un cadre bruité. L'avantage d'un tel choix est que les preuves théoriques de la convergence

4. Notons une erreur mineure dans [Bühlmann, 2006, Lutz and Bühlmann, 2006] où ce sont les résidus théoriques non observés $R_k = f - G_k$ qui sont utilisés pour choisir la succession des régresseurs φ_k et non les résidus apparents $Y - G_k$ qui sont les seules estimations disponibles pour l'algorithme.

Algorithm 5 Weak Greedy Algorithm (Cadre bruité)[Bühlmann, 2006]

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

Ensure: Paramètre de *shrinkage* $\nu \in]0,1]$, Itération maximale N_n

Prédicteur $G_0 = 0_H$ et Résidu $R_0 = f$.

$k \leftarrow 0$

while $k < N_n$ **do**

Choix d'un élément $\varphi_k \in \mathcal{D}$ suffisament corrélé avec le résidu « apparent » :

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

Mise à jour de la prédiction

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

et des résidus théoriques non observés

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

$k \leftarrow k + 1$

end while

(cadre déterministe) et consistance (cadre bruité) sont des adaptations simples des algorithmes univariés. Cependant, un tel choix ne prend finalement pas en compte l'amplitude des erreurs globales d'approximation, coordonnée par coordonnée. Ceci peut être un peu préjudiciable dans le cas bruité puisque l'on ne dispose pas d'une infinité d'itérations de Boosting pour régresser chaque coordonnée : l'itération maximale N_n est théoriquement dépendante de n et en pratique on stoppe l'algorithme par le biais d'un critère de type AIC et il est donc important de bien choisir l'ordre des coordonnées à prédire.

Algorithmes Boost-Boost multivariés déterministes Nous avons développé pour le contexte multivarié un algorithme qui va répartir l'effort de Boosting sur chaque coordonnée de H_m au fûr et à mesure des itérations afin d'éviter le défaut de l'approche multivariée précédemment évoquée. Nous proposons deux façons de sélectionner i_k et la méthode est décrite dans l'algorithme 6.

Nous procédons donc en deux temps à l'étape k : chercher d'abord la coordonnée i_k possédant le plus d'information à prédire, puis trouver le meilleur régresseur φ_k pour cette coordonnée. Il est alors possible de démontrer la convergence de l'algorithme basé sur la norme $\mathbb{L}^2$ des résidus. À nouveau, la convergence va dépendre de la taille de f et des paramètres de shrinkage μ, γ et ν qui appartiennent toutes à $(0, 1]$.

Théorème 2.3.2 (Algorithme Boost-Boost déterministe (norme $\mathbb{L}^2$ des résidus)) *Soit* $f = (f^1, \dots f^m) \in H_m$ *telle que toutes ses coordonnées* $f^j \in \mathcal{A}(\mathcal{D}, B)$. *Alors, pour tout* $k \geqslant m$, *l'algorithme 6 utilisant (2.9) converge : il existe une constante* $C_B > 0$ *ne dépendant que de* B *telle que*

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

La démonstration est une extension de la technique de [DeVore and Temlyakov, 1996], l'idée étant que lorsqu'on fait assez d'itérations, on sélectionne au moins une coordonnée un nombre

Algorithm 6 Algorithme Boost-Boost (Cadre déterministe)[19]

Require: Dictionnaire $\mathcal{D}$, fonction $f \in H$ à approcher.

Ensure: Paramètres de *shrinkage* $(\gamma, \nu, \mu) \in (0,1]^3$, Itération maximale N

Prédicteur $G_0 = 0_H$ et Résidu $R_0 = f$.

$k \leftarrow 0$

while $k < N$ **do**

Choix de la coordonnée i_k à booster

$$\|R_k(f^{i_k})\|^2 \geqslant \mu \max_{1 \leqslant i \leqslant m} \|R_k(f^i)\|^2 \qquad \text{[Norme } \mathbb{L}^2 \text{ des résidus]} \qquad (2.9)$$

ou bien

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

Choix du régresseur $\varphi_k \in \mathcal{D}$ suffisament corrélé avec R_k : $|\langle R_k, \varphi_k \rangle| \geqslant \nu \max_{g \in \mathcal{D}} |\langle R_k, g \rangle|$

Mise à jour de la prédiction

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

Mise à jour des résidus

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

$k \leftarrow k + 1$

end while

suffisant de fois et on exploite ensuite la procédure de sélection (2.9) pour déduire une décroissance globale des normes des résidus.

L'algorithme basé sur la sélection de la coordonnée i_k par le biais des sommes des corrélations (2.10) peut également être analysé en supposant un contrôle sur la cohérence du dictionnaire. On définit

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

et on suppose que chaque f^j est S parcimonieuse, c'est-à-dire

$$f^j = \sum_{i=1}^{p} a_i^j g_i \qquad \text{avec} \qquad \|a_\cdot^j\|_0 \leqslant S.$$

Il est alors possible de démontrer le résultat suivant.

Théorème 2.3.3 (Algorithme Boost-Boost déterministe (somme des corrélations à $\mathcal{D}$)) *Soit $f = (f^1, \dots f^m) \in H_m$ telle que toutes ses coordonnées $f^j \in \mathcal{A}(\mathcal{D}, B)$. Si chaque f^j est S parcimonieuse avec $\rho((1+\nu^{-1})S - 1) < 1$, alors, il existe une constante $C_{\rho,S,B}$ ne dépendant que de la cohérence du dictionnaire et de B telle que pour tout $k \geqslant 1$;*

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

Ce lien entre la cohérence de $\mathcal{D}$, l'indice de parcimonie et l'approximation de f par algorithmes de Boosting a déjà été faite par différents auteurs ([Temlyakov and Zheltov, 2011, Tropp, 2004] par exemple). En particulier, l'hypothèse $\rho(2S-1) < 1$ (obtenue pour le cas particulier $\nu = 1$) permet de démontrer que l'indice de parcimonie des résidus est une fonction décroissante tout au long des itérations du boosting et que l'algorithme n'injecte pas dans l'approximation de f des "mauvais" éléments du dictionnaire. Il n'est donc pas surprenant d'obtenir un résultat de ce type dans notre contexte.

2.3.3 Extension des algorithmes Boost-Boost multivariés aux situations bruitées

Il est possible de rendre compatibles les méthodes décrites dans l'algorithme 6 aux situations bruitées. En reprenant les notations du paragraphe 2.3.1 introduites pour les données empiriques, on peut décrire à nouveau l'adaptation de notre méthode Boost-Boost dans l'algorithme 7.

Notons qu'avec l'approche utilisant la norme $\mathbb{L}^2$ des résidus, il est impératif de ne pas utiliser de shrinkage sur μ, du moins d'un point de vue théorique. En reprenant l'idée de preuve de [Bühlmann, 2006] qui considère un algorithme de Boosting déterministe obtenu à partir de la succession de régresseurs aléatoirement sélectionnés par l'algorithme bruité, il est possible de montrer la consistance de l'algorithme 7. On peut supposer que le nombre de régresseur p_n dans $\mathcal{D}$ peut grandir avec n. Plusieurs hypothèses sont nécessaires pour obtenir la consistance statistique.

La première hypothèse est une hypothèse conjointe sur la structure du design et du dictionnaire $\mathcal{D}$. C'est une hypothèse technique permettant d'obtenir des résultats de loi des grands nombres "uniformes".

Hypothèse 1 ($\mathbf{H}_\mathcal{D}$) *Pour tout $g_j \in \mathcal{D}$, $g_j(X)$ possède un moment d'ordre 2 normalisé $\mathbb{E}[g_j(X)^2] = 1$ et est essentiellement bornée*

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

$$\sup_{1 \leqslant j \leqslant p_n, n \in \mathbb{N}} \|g_j(X)\|_\infty < \infty.$$

Algorithm 7 Algorithme Boost-Boost (Cadre bruité)[19]

Require: Dictionnaire $\mathcal{D}$, fonction $f \in H$ à approcher.

Ensure: Paramètres de *shrinkage* $(\nu, \gamma, \mu) \in (0,1]^2$, Itération maximale N_n

Prédicteur $\hat{G}_0 = 0_H$ et Résidu $R_0 = f$.

$k \leftarrow 0$

while $k < N$ **do**

Choix de la coordonnée i_k à booster

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

ou bien

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

Choix du régresseur $\varphi_k \in \mathcal{D}$ suffisament corrélé avec le résidu apparent $Y^{i_k} - \hat{G}_k^{i_k}$:

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

Mise à jour de la prédiction

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

Mise à jour des résidus théoriques non observés

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

$k \leftarrow k + 1$

end while

L'hypothèse suivante fixe le cadre statistique de la très grande dimension. Il est possible d'estimer quelque chose dès lors que $\log p << n$.

Hypothèse 2 ($\mathbf{H_{p_n}}$) *Le nombre de régresseurs p_n de $\mathcal{D}$ vérifie*

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

pour $\xi \in]0,1[$ et $0 < C < \infty$.

Enfin, l'hypothèse principale est la suivante et traduit une certaine parcimonie dans la décomposition sur $\mathcal{D}$ des éléments à estimer. Elle est évidemment vraie dès qu'on impose un indice de parcimonie fixe par rapport à n.

Hypothèse 3 ($\mathbf{H_f}$) *La fonction $f = (f^1, \dots f^m)$ à prédire se décompose grâce au dictionnaire $\mathcal{D}$*

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

et chaque coordonnée est S-parcimonieuse, avec S indépendant de n. La suite $(\gamma_i^{(j)})_{n \in \mathbb{N}, 1 \leqslant j \leqslant m, 1 \leqslant i \leqslant p_n}$ satisfait donc

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

L'hypothèse sur la nature du bruit permet d'appliquer des inégalités de concentration avec une méthode de troncature.

Hypothèse 4 ($\mathbf{H}_\epsilon$) *Les variables aléatoires $(\varepsilon_\ell)_{\ell=1\dots n}$ sont i.i.d centrées dans $\mathbb{R}^m$ de covariance Id_m indépendantes des $(X_\ell)_{\ell=1\dots n}$, et telles que*

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

pour $s > \frac{2}{\xi}$ où ξ est donné dans l'hypothèse 2 ($\mathbf{H}_{\mathbf{p_n}}$]).

Elle est satisfaite dès que les queues de distribution sont de nature Gaussienne ou Laplace par exemple.

Enfin, la dernière hypothèse sur l'amplitude des coefficients actifs n'est utile que pour obtenir un résultat de consistance sur l'estimation du support de f.

Hypothèse 5 ($\mathbf{H}_\mathcal{S}$) *Les coefficients actifs dans la décomposition S parcimonieuse de chaque coordonnée f^i satisfont :*

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

avec $\kappa < 1/2$.

On démontre alors le premier résultat suivant sur la reconstruction en support des algorithmes Boost-Boost.

Théorème 2.3.4 (Consistance en support des algorithmes de Boost-Boost) *Les trois points suivants sont satisfaits avec grande probabilité :*

i) *Supposons que les hypothèses 1-4 sont satisfaites ($(\mathbf{H}_\mathcal{D}), (\mathbf{H}_{\mathbf{p_n}}), (\mathbf{H}_\mathbf{f}), (\mathbf{H}_\epsilon)$) et que chaque composante f^j est S parcimonieuse vérifiant $\rho((1+\nu^{-1})S-1) < 1$. Alors il existe une valeur explicite du paramètre de Shrinkage γ^* tel que pour tout $\gamma \in (0,\gamma^*)$ l'algorithme Boost-Boost (norme $\mathbb{L}^2$ des résidus) ne sélectionne que des "bons" coefficients dès lors qu'on effectue un nombre $(k_n)_{n\in\mathbb{N}}$ d'itérations (avec grande probabilité).*

ii) *Si on suppose en plus que l'hypothèse (5) $\mathbf{H}_\mathcal{S}$ est vraie, alors il existe une valeur maximale $\kappa^*(\gamma, S)$ explicite qui assure qu'on estime bien le support de chaque coordonnée (avec grande probabilité) dès lors que $\kappa \leqslant \kappa^*(\gamma, S)$.*

iii) *Si par ailleurs $p_n = \underset{n\to+\infty}{o}(\sqrt{n})$, il en est de même pour l'algorithme de Boost-Boost basé sur la somme des corrélations à $\mathcal{D}$.*

Ce précédent résultat de consistance en support est nouveau pour les algorithmes WGA, et était déjà connu pour d'autres méthodes d'estimation parcimonieuse. La barre $n^{-\xi/2}$ correspond au seuil en deçà duquel il n'est pas possible de détecter de manière fiable un coefficient allumé dans la décomposition des f^i. Si cette hypothèse n'est pas satisfaite, on obtient par ailleurs uniquement un résultat de stabilité : l'algorithme ne sélectionne que des bons coefficient avec grande probabilité. Ces algorithmes possèdent donc des propriétés légèrement supérieures, d'un point de vue théorique à d'autres méthodes parcimonieuses où peu de choses sont connues dès que la condition sur l'amplitude n'est pas satisfaite. Ce qui permet d'obtenir un tel résultat ici est le paramètre de shrinkage γ, dont la valeur ne doit excéder théoriquement 13/18 par exemple dans le cas de l'algorithme Boost-Boost basé sur la norme $\mathbb{L}^2$ des résidus.

L'utilisation de ce premier résultat permet également d'obtenir le théorème suivant, à noter que l'hypothèse (5) $\mathbf{H}_\mathcal{S}$ n'est pas nécessaire pour l'obtention de telles consistances.

Théorème 2.3.5 (Consistance des algorithmes de Boost-Boost) *Supposons que les hypothèses 1-4 sont satisfaites ($(\mathbf{H}_{\mathcal{D}}), (\mathbf{H}_{\mathbf{p_n}}), (\mathbf{H}_{\mathbf{f}}), (\mathbf{H}_{\epsilon})$) et que chaque composante f^j est S parcimonieuse vérifiant $\rho((1+\nu^{-1})S-1) < 1$, alors, pour $\gamma < \gamma^*$ il existe $(k_n)_{n\in\mathbb{N}}$ une suite croissant suffisament lentement vers $+\infty$ telle que l'algorithme Boost-Boost basé sur la norme $\mathbb{L}^2$ des résidus vérifie*

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

Si on suppose en plus que $p_n = \underset{n\to+\infty}{o}(\sqrt{n})$ alors, le résultat reste vrai pour l'algorithme Boost-Boost basé sur la somme des corrélations à $\mathcal{D}$ qui vérifie

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

Le nombre de variables peut donc grandir exponentiellement vite avec le nombre d'observations n, ce qui correspond au bon ordre $\log p_n \sim Cn$ qui est connu dans les travaux en régression parcimonieuse. L'outil statistique principal est une loi des grands nombres uniforme associée à l'hypothèse $(\mathbf{H}_{\epsilon})$. Remarquons tout de même que k_n varie logarithmiquement avec n (un ordre identique à ceux obtenus dans [Bühlmann, 2006]) et ce résultat est plus faible que les résultats qui sont généralement obtenus en régression pénalisée parcimonieuse comme le Lasso. Nous renvoyons à [19] pour les éléments techniques des preuves de ces théorèmes.

2.3.4 Résultats numériques

Nous décrivons ici brièvement certains résultats obtenus *via* les algorithmes de Boosting et renvoyons à [19] ou [20] pour de plus amples simulations.

La première simulation consiste en l'étude d'un jeu de données déjà utilisé dans [Lutz and Bühlmann, 2006]. On observe la matrice de réponse Y de taille $n \times m$ alors que les observations X sont de taille $n \times p$ et que le modèle s'écrit $Y = X\theta + \epsilon$ où ϵ est un bruit gaussien $\mathcal{N}(0, I_n)$. Par ailleurs, des corrélations sont introduites dans le jeu de données et chaque paire de variables (g_j, g_k) (pour $1 \leqslant j, k \leqslant p$) qui sont telles que $\rho(g_j, g_k) = 0.9^{|k-j|}$. Chaque colonne de θ sera s-parcimonieuse.

La seconde simulation étudie plus précisément le cas de l'inférence d'un réseau de régulation de p gènes : chaque niveau d'expression (codé dans les n observations E_i) d'un gène $E_{i,j}$, pour $1 \leqslant j \leqslant p$ est implicitement régulé par le niveau d'expression des autres gènes $E_{i,k}, k \neq j$, ce qui se traduit par une modélisation du réseau en $E = E\theta + \epsilon$ avec θ une matrice à estimer à diagonale nulle pour éviter une régression triviale et insignifiante pour notre problème. Là encore, ϵ est un bruit gaussien $\mathcal{N}(0, I_n)$.

La dernière simulation reprend le principe de la première en introduisant des paramètres de parcimonie s qui peuvent être différents dans les colonnes de la matrice θ, et en introduisant des corrélations plus fortes (± 0.9) entre les variables explicatives que ce qui est fait dans la première expérience.

Toutes les figures précédentes proposent l'évolution de la précision des algorithmes en fonction de leurs puissances. Il s'agit donc en abscisse de la proportion de coefficients retrouvés par les algorithmes et en ordonnée de la proportion de bonnes prédictions de coefficients allumés. Ces courbes présentent donc des résultats 'en support' et non en terme d'erreur de prédiction (on consultera [19] pour des résultats complémentaires).

Les deux premiers jeux de données nous prouvent que la plupart du temps, les méthodes de boosting sont relativement comparables et se comportent favorablement par rapport aux méthodes classiques de Random Forest ou Bootstrap Lasso. En général, nos études numériques laissent à penser que les algorithmes de prédiction par réseaux bayésiens sont un peu moins

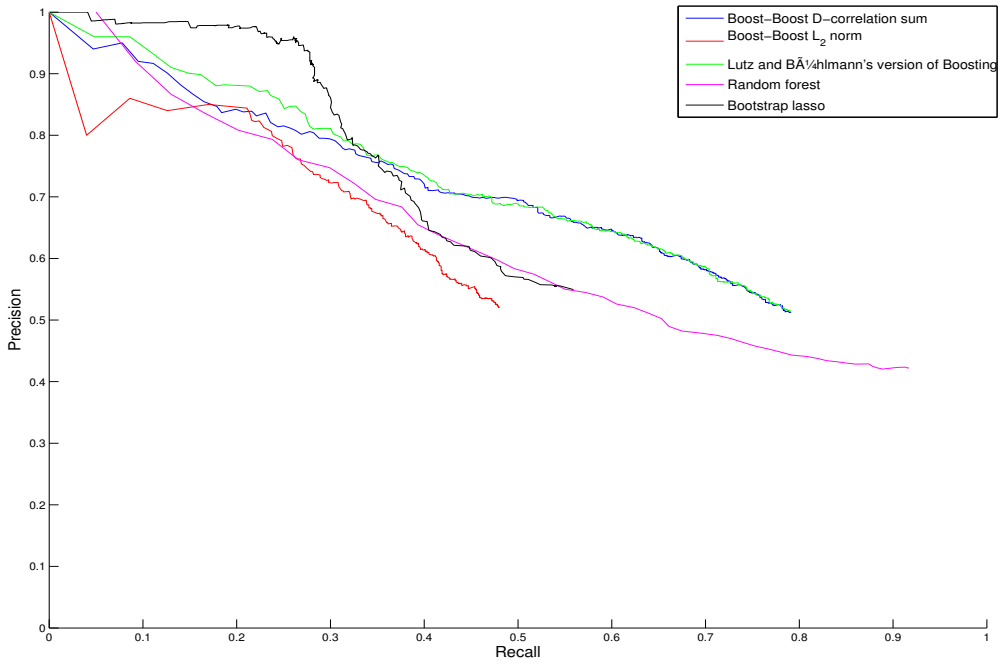


FIGURE 2.4 – Résultat de 5 méthodes de régression parcimonieuse sur le modèle jouet de [Lutz and Bühlmann, 2006]. Ici $p = 10$, $n = 50$, $m = 4$ et l'indice de parcimonie s vaut 2. En abscisse : la puissance, en ordonnée la précision.

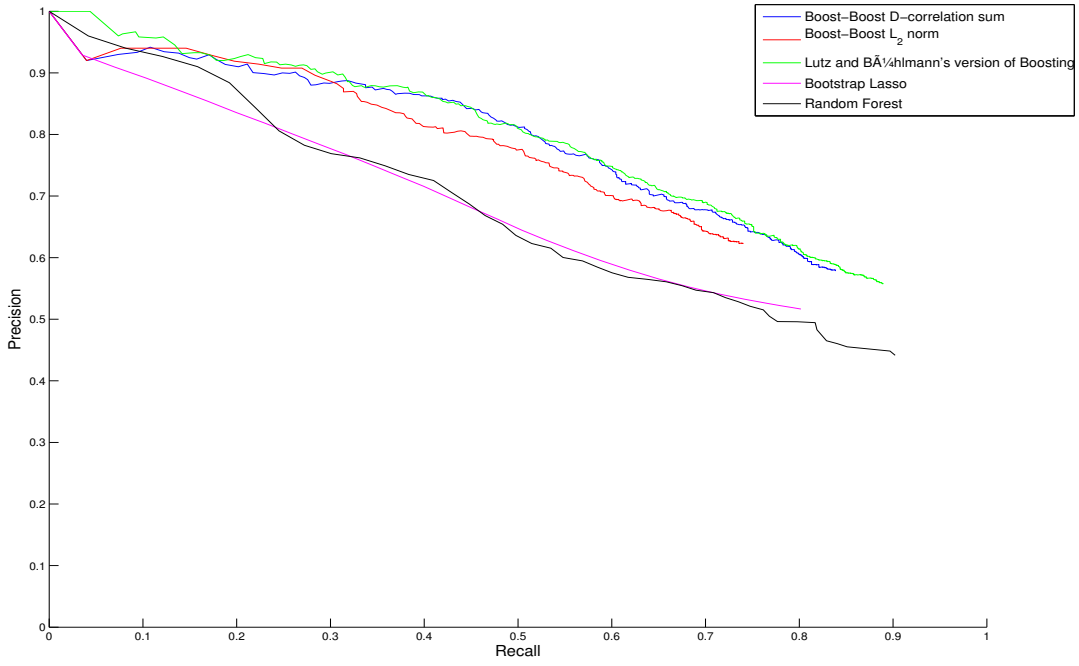


FIGURE 2.5 – Résultat de 5 méthodes de régression parcimonieuse sur le modèle d'inférence de réseau de régulation. Ici $p = 10$ et $n = 50$. En abscisse : la puissance, en ordonnée la précision.

performants (non illustré ici mais disponible dans [19] ou [20]). Notons que le premier jeu de données se place dans un cadre qui n'est pas réellement de la 'grande dimension' puisque dans

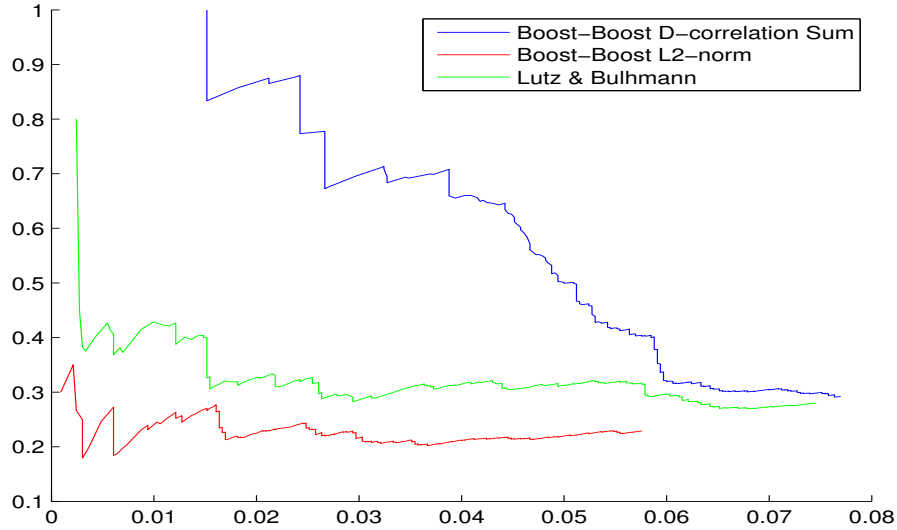


FIGURE 2.6 – Résultat des 3 méthodes de boosting sur le troisème modèle de régression avec variables fortement corrélées avec $m = 4$. Ici $p = 250$, $n = 50$ tandis que $s = (30, 100, 100, 100)$. En abscisse : la puissance, en ordonnée la précision.

ce cas, la dimension est de 40 alors que l'on observe 50 individus.

Dans le cas extrême où des corrélations de l'ordre de $\pm\frac{9}{10}$ sont présentes dans le jeu de données (Figure 2.6), on constate que l'algorithme de boosting utilisant la somme des corrélations au dictionnaire est sensiblement plus performant que les 2 autres algorithmes de boosting et étale mieux l'effort des premières itérations de boosting que l'approche de [Lutz and Bühlmann, 2006]. Par ailleurs, on constatera que l'algorithme maintient une précision acceptable jusqu'à une puissance de l'ordre de 5/100, mais ceci est obtenu dans un cadre de très grande dimension puisque $n = 50$ alors que la dimension du problème est 1000 ici.

2.3.5 Élargissements

Même si les résultats théoriques semblent tout à fait satisfaisants, les performances numériques sont un petit peu décevantes dès lors qu'on atteint le régime où $n << p$ et que l'indice de parcimonie n'est plus petit (illustré par la courbe 2.6). Ce phénomène est prévisible grâce aux résultats théoriques précédents en terme d'équilibre entre S, p, n, γ et κ (voir Théorème 2.3.4). Il serait relativement légitime de s'intéresser à une famille de méthodes totalement ignorée dans ces simulations que sont les méthodes d'inférence Bayésienne. Un point d'entrée intéressant dans le contexte de la grande dimension serait d'utiliser les travaux récents de [Castillo and van der Vaart A., 2012].

Chapitre 3

Statistiques de modèles déformables

Dans ce chapitre, je propose une synthèse des problèmes que j'ai étudiés et des résultats obtenus sur le thème des statistiques et modèles déformables. L'objet de notre étude est de décrire des méthodes d'estimation pour les signaux et images, nous nous plaçons donc naturellement dans un cadre fonctionnel et les objets qu'on cherche à reconstruire appartiennent à un espace $\mathcal{H}$ abstrait pour lequel on fournira plus de détails plus loin.

3.1 Modélisation d'action de déformations

Tous les problèmes d'estimation présentent ici le point commun de présenter des observations qui chacune a subie une perturbation par un double aléas : une déformation aléatoire d'une même forme « moyenne » $f^\star$ pour toute les observations et ensuite un bruit de mesure additif. Ainsi, chacune des n observations est décrite par une équation

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

où ε_i désigne le bruit de mesure, f_i est la forme déformée (aléatoire) appartenant à $\mathcal{H}$ et Y_i l'observation réellement obtenue. Si $\mathcal{H}$ est un ensemble de fonctions définies sur Ω, on peut définir chaque f_i en nous inspirant des idées de [Grenander, 1993b]. Ainsi, chaque f_i est définie au travers de

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

où Z_i désigne la variation en amplitude et g_i désigne l'action de déformation. Ainsi, si l'action de Z_i est linéaire, ce n'est par contre pas le cas pour l'action de g_i qui traduit une injection de Ω dans Ω. Dans tout ce qui suit, on ne traitera que des problèmes d'estimation lorsque $Z_i = 0$ puisque nous n'avons pas considéré dans nos travaux les variations photométriques ou en amplitude des signaux.

3.1.1 Déformation rigide

Il existe schématiquement 2 classes de déformations homéomorphes de Ω, les rigides et les élastiques. Les déformations rigides sont les plus simples et correspondent à l'action d'un groupe de Lie de dimension finie sur Ω. L'exemple typique est celui où par exemple $\Omega = \mathbb{R}^d$ et le groupe de déformation est le groupe des translations et dans ce cas les observations sont simplement données par

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

où τ_i est le paramètre de translation aléatoire. Ici, l'action de G est tout simplement $g.x = x - g$ pour tout x de Ω.

Bien entendu, cette situation peut englober des situations plus complexes en dimension supérieure lorsque G comprend des rotations, translations, homothéties, ...

3.1.2 Déformation élastique

La seconde classe de déformation bijective est nettement plus complexe à définir mais permet de proposer un cadre « élastique »aux déformations de Ω. Ces modèles de déformations ont été introduites *via* les flots d'équations différentielles par [Miller and Younes, 2001, Trouvé and Younes, 2005].

Afin de modéliser des déformations de Ω qui garantissent la bijectivité, l'idée est la suivante : si v_i désigne une application de $\mathcal{C}(\Omega,\Omega)$, et si ϵ est un réel positif assez petit, alors l'application $\phi_1 = \mathrm{Id} + \epsilon v_i$ est toujours homéomorphe (sur son image). Ainsi, $\phi_p \circ \phi_{p-1} \circ \cdots \circ \phi_1$ est également homéomorphe. Finalemement, en remarquant que

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

on constate que le modèle de composition de petites déformations peut être généralisé. Si on se donne une famille d'applications $(v_t)_{t\in[0;1]}$ de $\mathcal{C}(\Omega,\Omega)$, et qu'on considère la famille de fonctions

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

alors (3.3) admet une unique solution $(\phi_t)_{t\in[0;1]}$ initialisée en $\phi_0 = \mathrm{Id}_\Omega$ dès que $\int_0^1 \|v_s\| ds < +\infty$. Par ailleurs, pour tout temps t, ϕ_t est un homéomorphisme de Ω dans $\phi_t(\Omega)$. Enfin, pour imposer que les homéomorphismes soient surjectifs dans Ω, il est suffisant d'imposer qu'ils correspondent à l'identité sur le bord de Ω, ce qui est vrai dès que $\forall t \in [0;1], \forall x \in \partial\Omega v_t(x) = 0$.

Ainsi, nous avons à notre disposition deux façons de générer des observations f_i déformées d'une observation initiale $f^\star$. Bien entendu, il est parfois plus réaliste mais plus compliqué théoriquement de manipuler des déformations élastiques.

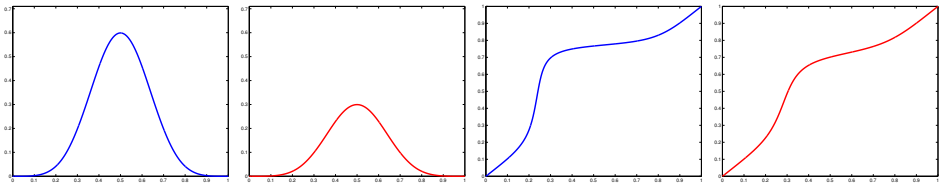


FIGURE 3.1 – Exemple uni-dimensionnel de deux homéomorphismes ϕ_1 de $[0;1]$ (à droite) générés par des champs de vecteur v homogènes en temps (à gauche).

3.1.3 Régression sous contrainte de monotonie

Mon premier travail dans le contexte des problèmes d'estimation avec des objets déformables à consister à exploiter la nature très spécifique des homéomorphismes en dimension 1 qui sont naturellement des fonctions monotones. Nous exploitons dans [10] cette simple remarque pour construire une méthode de régression sous contrainte de monotonie. Ce lien est décrit dans le résultat suivant qui est la clef de la réussite de notre méthode d'estimation. Si I désigne un intervalle de $\mathbb{R}$, on définit $\mathcal{H}^m(I)$ l'espace de Sobolev par

$$\mathcal{H}^m(I) = \{f : I \to \mathbb{R}, f^{(m-1)} \text{ est continue sur I et } \int_I |f^{(m)}(x)|^2 dx < +\infty\},$$

on a alors la caractérisation suivante de fonctions monotones (sur $I = [0,1]$ par exemple).

Théorème 3.1.1 *Notons* $\tilde{\mathcal{H}} = \mathrm{Span}\{1, x\} + \mathcal{H}^m(\mathbb{R})$ *et* $m \geqslant 2$. *Pour tout* $f \in \mathcal{H}^m([0,1])$ *strictement croissante sur* $I = [0,1]$, *on définit*

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

Alors il existe un champ de vecteur $(v_t^f)_{t\in[0,1]}$ *tel que* $v_t^f \in \tilde{\mathcal{H}}, \forall t \in [0;1]$ *tel que*

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

De plus, pour tout $t \in [0,1]$, *on a*

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

On notera que l'objet à estimer f correspond donc à ϕ_1. L'idée est d'obtenir un estimateur monotone au travers de l'estimation de $(v_t)_{t\geqslant 0}$ puis utiliser (3.3) pour obtenir ϕ_1. Pour tout $t \in [0;1]$, on sait par le biais du théorème 3.1.1 que v_t « envoie » $tf(x) + (1-t)x$ sur $f(x) - x$.

D'un point de vue statistique, on dispose d'observations $(x_1, y_1), \dots (x_n, y_n)$ telles que

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

où les variables $(\epsilon_i)_{i\in\{1\dots n\}}$ sont centrées de variance σ^2. On cherche un estimateur monotone $\hat{f}_n$ tel que son risque quadratique défini par

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

soit faible et qui soit de plus monotone. L'idée est la suivante : en choisissant $\hat{f}_n^0$ un estimateur non contraint de f, on construit alors un estimateur contraint $\hat{f}_n^c$ monotone et qui hérite des mêmes propriétés asymptotiques que $\hat{f}_n^0$. Cette étape substitue donc la phase de projection sur l'espace des contraintes des approches classiques. Pour cela, on cherche le champ de vecteur $v^{n,\lambda} = (v_t^{n,\lambda})_{t\in I}$ qui estime $(v_t)_{t\geqslant 0}$ tel que pour tout $t \in [0;1]$, $v_t^{n,\lambda}$ appartient à $\tilde{\mathcal{H}}$ s'écrit

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

On munit $\tilde{\mathcal{H}}$ d'une structure d'espace hilbertien à noyaux reproduisant (R.K.H.S) par le biais d'un noyau K, et une manière de trouver $v_t^{n,\lambda}$ est de chercher la solution du problème d'optimisation

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

Il est alors possible (voir théorème 5.1 de [10]) d'obtenir sous un certain nombre d'hypothèses techniques sur le noyau K et le coefficient de pénalisation λ_n utilisé qu'avec grande probabilité :

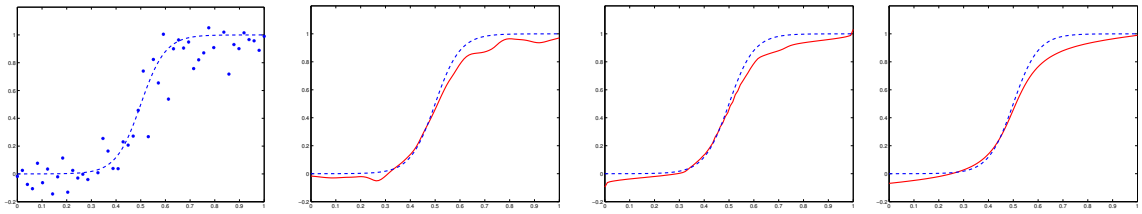
$$R(\hat{f}_n^c, f) \leqslant C(R(\hat{f}_n, f) + \lambda_n).$$

Ce résultat peut être étendu à l'étude du risque quadratique sur $[0;1]$ par le biais du théorème suivant.

Théorème 3.1.2 *Si* $f \in \mathcal{H}^m(I)$ *est telle que* $f' > 0$ *sur* I *et si l'on choisit* $\lambda_n = 1/n$, *alors* $\hat{f}_n^c$ *construit à partir de l'estimateur* $\hat{f}_n^0$ *défini dans [Speckman, 1985] non contraint est un estimateur monotone et est asymptotiquement optimal au sens minimax :*

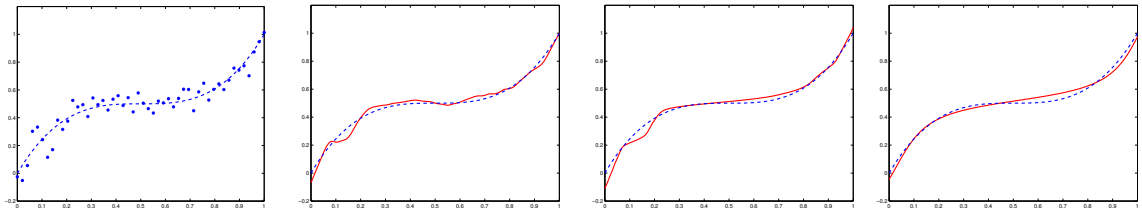
$$R_n(\hat{f}_n^c, f) = O(n^{-2m/(2m+1)}).$$

Les figures (3.2), (3.3) et (3.4) sont issues de l'étude expérimentale menée dans [10] et illustrent la monotonisation d'estimateur non contraint obtenue par le biais de notre flot de champ de vecteur. On peut également mentionner que les travaux décrits dans [10] peuvent s'étendre naturellement à la construction de difféomorphismes en dimension supérieure afin de mettre en correspondance deux jeux de *landmarks*.



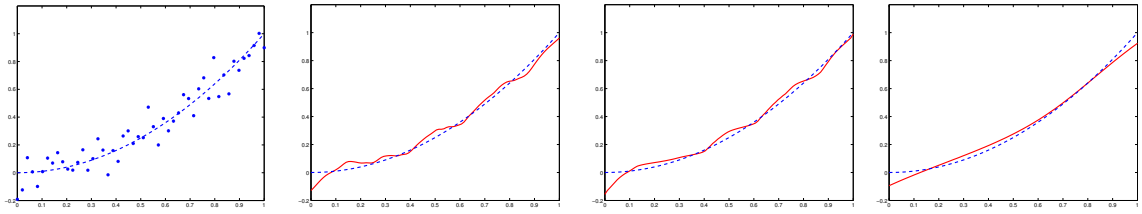
(a) (b) (c) (d)

FIGURE 3.2 – Signal m_1 : la ligne pointillée est l'inconnue f, (a) données brutes avec SNR = 3, (b) Estimateur non contraint $\hat{f}_n^0$, (c) Estimateur de Dette et al., (d) Estimateur monotonisé $\hat{f}_n^c$ à partir de $\hat{f}_n^0$.



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

FIGURE 3.3 – Signal m_2 : la ligne pointillée est l'inconnue f, (a) données brutes avec SNR = 3, (b) Estimateur non contraint $\hat{f}_n^0$, (c)Estimateur de Dette et al., (d) Estimateur monotonisé $\hat{f}_n^c$ à partir de $\hat{f}_n^0$.



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

FIGURE 3.4 – Signal m_3 : la ligne pointillée est l'inconnue f, (a) données brutes avec SNR = 3, (b) Estimateur non contraint $\hat{f}_n^0$, (c)Estimateur de Dette et al., (d) Estimateur monotonisé $\hat{f}_n^c$ à partir de $\hat{f}_n^0$.

3.2 Modèle déformé et bruit blanc, loi des déformations connue

Dans cette section, je décris mes travaux portant sur l'estimation de f dans le modèle de déformation lorsque les observations (3.1) sont données par un modèle de bruit blanc gaussien et que les f_i sont définies par (3.2). L'objectif est une description précise de méthode d'estimation lorsque le nombre n d'observations tend vers $+\infty$. Nous nous plaçons dans le contexte où la loi des déformations notées g est connue. Cette hypothèse est fondamentale pour toute la section.

3.2.1 Modèle de courbes translatées aléatoirement

Je consacrerai plus de temps à décrire cette partie (qui est la plus simple techniquement) qu'à ses généralisations (modèle plus riche de déformations, bruit poissonien) car elle contient déjà toutes les idées principales des autres situations considérées dans [8] et [17].

Le cas le plus simple consiste à envisager les f_i comme issues d'une simple translation aléatoire d'une fonction f inconnue périodique de période 1 *dans le cas où la loi* g *des translations est connue*. Ainsi, on considère le problème d'estimation de f dans le modèle de bruit blanc

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

Approche par déconvolution Le modèle (3.6) peut être compris en considérant la base de Fourier $(e_k)_{k\in\mathbb{Z}}$ sur laquelle chaque mouvement brownien $W_i(x)$ se décompose en coefficients gaussiens indépendants :

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

où les $(\epsilon_{j,k})_{j,k}$ sont des variables aléatoires i.i.d. $\mathcal{N}(0,1)$.

L'idée pour la construction d'un estimateur de f est qu'en connaissant les coefficients de Fourier $(\gamma_k)_{k\in\mathbb{Z}}$ de g, on peut estimer ceux de f en remarquant que

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

En effet, l'égalité précédente est vraie en espérance dès lors que $\gamma_k \neq 0$ et finalement la loi des grands nombres permet d'espérer une bonne estimation de f. Il convient néanmoins de remarquer que l'inversion de γ_k peut devenir hasardeuse lorsque k est grand puisque $\gamma_k \mapsto 0$ lors $k \to \pm\infty$. On rencontre ici un phénomène bien connu dans le contexte des problèmes inverses lorsque l'opérateur à inverser est mal conditionné. Ce phénomène est tout à fait logique pour notre méthode d'estimation puisque nous optons pour une méthode de déconvolution et que les γ_k correspondent aux valeurs propres de l'opérateur de convolution que l'on cherche à inverser.

Notons tout de même qu'il ne semble pas évident que cet aspect « problème inverse » soit naturel à la vue du modèle (3.6) mais peut être propre à notre façon d'envisager l'estimation (3.7).

Estimateur par seuillage et vitesse de reconstruction dans les espaces de Besov Pour ce modèle, nous mesurons la vitesse de reconstruction en utilisant le risque quadratique moyen : pour tout estimateur $\hat{f}$, la perte est définie par

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

Il est possible de construire $\hat{f}_n$ à partir d'une décomposition dans une base d'ondelettes. Plus précisément, si $(\psi_{j,k})_{j,k}$ et $(\phi_{j,k})_{j,k}$ désignent les fonctions d'échelle et d'ondelette de Meyer à l'échelle j et localisation k, on va chercher

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

où $\hat{c}$ et $\hat{\beta}$ seront les coefficients à estimer en fonction des observations. La description complète de cette méthode d'estimation est un petit peu technique et nous renvoyons au travail [9] pour plus de détails. L'idée principale sous-jacente à cette méthode est classiquement de limiter la taille de j_0 et j_1 en fonction de n et ν la régularité de g, et garder parmi tous les coefficients d'ondelettes calculés ceux dont l'amplitude est supérieure à un seuil calculé sur les données.

Le théorème de reconstruction obtenu dans [9] (dans un cadre fonctionnel un petit peu simplifié ici) peut s'énoncer ainsi.

Théorème 3.2.1 *Soient $f \in B^s_{2,2}$ inconnue, de régularité s inconnue, et g dont les coefficients de Fourier connus vérifient*

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

Alors l'estimateur $\hat{f}^H_n$ défini dans [9] vérifie

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

L'estimateur obtenu est un estimateur de type seuillage dur et est *adaptatif* en la régularité de la fonction s, et que cette adaptativité a été obtenue par le biais de l'utilisation de la base d'ondelette. Il faut noter que la méthode que nous proposons exploite la propriété fondamentale de la base de Meyer d'être à support borné en fréquence, ce qui permet à partir des coefficients de Fourier estimés d'estimer les coefficients d'ondelettes et ainsi déduire une estimation de la fonction f.

Remarquons enfin que la vitesse $n^{\frac{-2s}{2s+2\nu+1}}$ obtenue dans le théorème 3.2.1 est classique lorsqu'on étudie le problème de déconvolution direct par la loi g décrite dans le modèle

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

Il n'est ici pas surprenant de retrouver la même vitesse de reconstruction puisque l'inversion opérée dans (3.7) et qui est utilisée ensuite dans le calcul des coefficients d'ondelette de Meyer est utilisée également dans les approches de déconvolution.

Taux minimax de reconstruction Classiquement en statistique non paramétrique, le problème de la pertinence de l'estimation est abordé en déterminant une borne inférieure de reconstruction qu'on espère proche de la borne supérieure d'estimation. Ceci est décrit par la vitesse minimax définie à partir d'un nombre d'observations n et d'une classe de fonction $\mathcal{F}$ parmi laquelle on cherche l'estimateur. On note

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

$\hat{f}_n$ parcourant l'ensemble des estimateurs. Il faut donc comprendre $\hat{f}_n$ comme étant une fonction mesurable des n observations et $\mathcal{R}_n(\mathcal{F})$ représente alors la meilleure vitesse qu'on peut atteindre pour la pire fonction à estimer dans la classe $\mathcal{F}$.

Habituellement, on aborde la question de la borne inférieure en étudiant précisément le rapport de vraisemblance de deux hypothèses qui sont éloignées en norme L^2 et qui pourtant sont proches d'un point de vue statistique. On peut citer trois résultats qui donnent des minorations de risque utilisant ce formalisme : à savoir le lemme de Fano (voir [Ibragimov and Has′minskiĭ, 1981] par exemple) le lemme d'Assouad (utilisé ici) et la méthode de [Le Cam, 1986]. Dans [9], nous abordons la minoration du risque $\mathcal{R}_n(\mathcal{F})$ lorsque $\mathcal{F}$ est l'espace de Besov $B^s_{2,2}$ par le biais du lemme d'Assouad dont je rappelle une des variantes ici. On consultera [Tsybakov, 2003] pour un exposé exhaustif sur les variations autour de ce lemme.

Lemme 3.2.1 (Lemme d'Assouad ([Bretagnolle and Huber, 1979])) *Étant donné un réseau de fonctions $(f_\theta)_{\theta\in\Theta}$ formant un cube $\Theta = \{\theta = (\theta_1, \dots \theta_d) \in \{\pm 1\}^d\}$, on note le rapport de vraisemblance des n observations $\Lambda(f_{\theta'}, f_\theta)$. Si pour toute paire d'hypothèse $f_\theta, f_{\theta'}$ telle que $\|\theta - \theta'\|_0 = 1$, on peut minorer en probabilité le rapport de vraisemblance*

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

pour un certain $\beta > 0$ et $\alpha \in (0,1)$, alors

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

Ce lemme traduit effectivement la difficulté en perte quadratique d'estimer le bon f_θ lorsque les rapports de vraisemblance sont proches ($\beta > 0$) avec une probabilité suffisament éloignée de 0.

Si l'utilisation de ce lemme afin d'obtenir une minoration du risque est aisée dans le cas de la convolution directe de f par g, le travail à effectuer pour faire fonctionner cette approche est nettement plus complexe dans le cas du modèle (3.6). Nous donnons tout d'abord dans [9] un sens au rapport de vraisemblance de deux hypothèses dans le cadre du modèle (3.6) en utilisant un conditionnement à la valeur de chaque translation, et en utilisant le changement de mesure de Girsanov. Pour toute fonction mesurable ψ des données Y, on peut écrire

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

L'astuce consistant à passer par l'hypothèse nulle f_0 permet de remarquer que dans le cas où les données suivent une loi issue de $f = 0$ dans le modèle (3.6), la loi de Y ne dépend pas des décalages α et donc

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

Le changement de mesure entre deux hypothèses est alors défini par la relation

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

Nous démontrons dans [9] une minoration de $\Lambda(f_\theta, f_{\theta'})$ en probabilité pour un certain réseau de fonctions dans $B^s_{2,2}$ et obtenons alors le résultat qui suit.

Théorème 3.2.2 *Soit $A > 0$, et g vérifiant la propriété de décroissance (3.8), si $\nu > 1/2$ et $s > 2\nu + 1$ alors il existe C une constante C dépendant de A et s telle que*

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

Ce résultat démontre alors que la méthode d'estimation obtenue par problème inverse est asymptotiquement optimale au sens du risque minimax puisque les bornes supérieures et inférieures du risque coïncident à un facteur logarithmique près. Ainsi, on peut traduire le modèle (3.6) comme un problème inverse avec opérateur connu (issu de la loi des translations de loi g) mais bruité. La minoration en probabilité fait appel à des développements limités assez précis au deuxième et troisième ordre de ce rapport de vraisemblance, ainsi qu'à des résultats de concentration comme l'inégalité de Bernstein.

Remarquons que l'utilisation du lemme de Fano plutôt que de la méthode d'Assouad ne semble pas plus directe pour établir le résultat. En effet, en partant de (3.9), la divergence de Kullback-Leibler entre deux hypothèses f_θ et $f_{\theta'}$ s'écrit

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

Notons que dans la situation de convolution directe par la loi g décrite dans le modèle

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

la divergence de Kullback s'écrit

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

où

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

On constate alors que la simplification « Logarithme - Exponentielle » de la situation standard est impossible pour la vraisemblance donnée par (3.9). Il est uniquement possible d'appliquer l'inégalité de Jensen pour obtenir

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

et traiter un tel terme revient aux mêmes calculs que ceux qui sont faits en considérant l'aménagement du lemme d'Assouad.

Remarque 3.2.1 *Notons enfin que si la méthode d'obtention de l'estimateur est assez classique, ce n'est pas tout à fait le cas pour la minoration du risque. La clef se situe dans l'identification d'éléments de $\mathcal{F}$ tels que la loi des observations Y est indépendante des variables aléatoires non observées (ici les translations aléatoires). Une telle stratégie pourrait sans doute permettre d'aborder d'autres calculs de bornes inférieures en procédant donc à la même stratégie d'identification d'hypothèses « invariantes » aux variables aléatoires non observées.*

3.2.2 Action aléatoire pour l'estimation dans des groupes de Lie

Il est possible de considérer une généralisation du modèle (3.6) lorsque les déformations tirées aléatoirement sont des éléments modélisant des transformations comme des translations, des symétries en dimension 2 ou 3. Ceci peut avoir un intérêt lorsqu'on considère des problèmes de traitement d'imagerie médicale acquise par exemple par le biais d'une transformée de Radon, ou pour des problèmes d'imagerie en robotique (imagerie satellitaire) lorsqu'une photo est prise par un robot (satellite) dont l'objectif pourrait avoir subi des petites rotations/translations par rapport à sa position théorique.

Dans [8] [1], nous formulons un modèle sensiblement équivalent à (3.6). Étant donné un groupe de Lie G de transformations, *compact et semi-simple*, on s'intéresse à l'estimation de $f \in \mathbb{L}^2(G)$ (espace des fonctions complexes définies sur G, de carré intégrable pour la mesure de Haar) au travers des observations

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

Ici, h est la loi des déformations supposée connue des déformations, alors que celles-ci agissent sur G par translation à gauche et que f est toujours le signal à reconstruire. Tous les ingrédients de résolution pour parvenir à estimer f sont relativement contenus dans l'approche développée dans le paragraphé précédent. En utilisant le théorème de Peter-Weyl, les éléments de $\mathbb{L}^2(G)$ sont développables en série de Fourier sur les représentations irréductibles de G qui sont dénombrables du fait de sa compacité. Ainsi, la formule d'estimation de f s'appuie sur la reconstruction $\mathbb{L}^2$

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

où $\hat{G}$ désigne l'ensemble des représentations irréductibles de G, d_π est la dimension de la réprésentation π, et $c_\pi(f)$ est la matrice jouant le rôle de coefficients de Fourier de f sur π qui est un vecteur propre du Laplacien sur G de valeur propre λ_π.

Estimer f revient alors à trouver une façon d'estimer ses coefficients de Fourier pour les basses fréquences, et seuiller les hautes fréquences. Là encore, une hypothèse sur la régularité ν de h, et lorsque la décroissance s des coefficients de Fourier de f est connue, permet de donner un estimateur convergent de f. Si on note

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

on construit alors l'estimateur $\hat{f}_n^T$ par seuillage en enlevant certaines représentations π :

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

où $\hat{G}_T$ est l'ensemble des représentations dont la valeur propre associée λ_π est inférieure à T.

L'analyse harmonique précédente permet de plus de donner un cadre fonctionnel généralisant les espaces de Sobolev réels en considérant alors les espaces

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

Les propriétés classiques d'analyse harmonique permettent alors de conclure qu'il existe une bonne fréquence de coupure pour l'estimation lorsqu'on sait dans quel espace $H_s(A)$ appartient la cible (le paramètre principal étant finalement le s qui permet de bien seuiller dans ces espaces fonctionnels).

Théorème 3.2.3 *Supposons* h *connue de régularité* ν *et* $f \in H_s(A)$ *où* s *est connu et* $s > \dim(G)/2$. *Alors pour le choix* $T_n = n^{\frac{2}{2s+2\nu+\dim G}}$, *il existe* $K_1 \geqslant 0$ *tel que*

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

1. Ce travail est le résultat du stage de M2R de Claire Christophe effectué en 2011.

Par ailleurs, il est également possible de retravailler l'approche basée sur un rapport de vraisemblance similaire à celui donnée par (3.9) et obtenir une borne inférieure de reconstruction.

Théorème 3.2.4 *Supposons* h *de régularité* ν *et* $s > 2\nu + \dim(G)$, *alors il existe* $K_2 \geqslant$ *tel que*

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

Remarque 3.2.2 *Remarquons que le résultat du théorème 3.2.3 est plus faible que celui donné par le théorème 3.2.1. En effet, nous obtenons ici un estimateur non adaptatif en la régularité* s *de la fonction cible, alors que c'était le cas pour l'estimateur donné dans le théorème 3.2.1. Ceci est expliqué par l'utilisation d'une analyse de Fourier et d'un seuillage brutal alors que nous utilisions auparavant une décomposition en ondelettes et une méthode de seuillage dur des coefficients d'ondelettes. Il serait possible de parvenir à construire un estimateur adaptatif dans le cadre de l'estimation (3.10) en utilisant par exemple la méthode de [Lepski, 1991] qui a une implémentation simple mais un coût de calcul relativement important.*

3.2.3 Approche à horizon fini

Les deux précédentes sections apportent des réponses asymptotiques dans les modèles de déformation aléatoire. Il est possible de raisonner à nombre de courbes n fixées, dans [12], nous étudions à nouveau le modèle :

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

et étudions le risque quadratique $\mathbb{L}^2$ d'estimation en utilisant à nouveau l'analyse de Fourier. À k fixé on procède à une estimation préliminaire du coefficients de Fourier $c_k(f)$ en utilisant la moyenne des $(\theta_{j,k})_{j=1\dots n}$. On utilise alors une technique de filtrage par un jeu de coefficients $(\lambda_k)_{k\in\mathbb{Z}}$ pour calculer l'estimateur $\hat{\theta}(\lambda)$. Plus précisément, on pose

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

Le risque quadratique d'estimation de f par $f_{\hat{\theta}(\lambda)}$ se décompose alors en

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

Le terme de biais est standard et on constate que le terme de variance est composé de deux termes, le premier également standard dû au modèle de bruit blanc en problème inverse tandis que le second correspond à la division par γ_k dans (3.7) et non $\tilde{\gamma}_k = \frac{1}{n}\sum_{j=1}^{n} e^{-i2\pi k\tau_j}$.

Bien entendu, $|c_k(f)|^2$ et donc R sont inconnus dans la formule précédente et nous ne pouvons rechercher en pratique le meilleur filtrage λ pour estimer concrètement f. Par contre, il est possible de construire une estimation de $|\hat{\Theta}_k|^2$ de $|c_k(f)|^2$ et utiliser alors l'approche URE (*Unbiased Risk Estimation*) dans notre contexte. On définit pour $\alpha \in [0;1]$

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

En considérant une sous-classe de filtre symétriques et monotones en $|k|$:

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

on construit l'estimateur empirique

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

Il est alors possible de démontrer que l'estimateur $\hat{\theta}(\hat{\lambda}_\alpha)$ satisfait une « inégalité oracle ».

Théorème 3.2.5 *Si les coefficients de Fourier de* g *satisfont l'hypothèse de décroissance (3.8), alors il existe* $\gamma_1 \in (0,1)$ *tel que pour tout* $\gamma \in (0,\gamma_1)$,

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

où $h_{\gamma,n} \to 0$ *lorsque* $\gamma \to 0$ *et* $n \to +\infty$, *et* $\Gamma_{\gamma,n,\epsilon^2}(c(f),\alpha)$ *est une fonction explicite de* (γ, n, ϵ^2) *et* $(c(f),\alpha)$.

La description de la fonction Γ précédente est relativement technique et on consultera [12] pour plus de détails. Les termes dans la fonction Γ sont principalement de deux sortes : un avec une décroissance en ϵ^2/n et l'autre en $\log^2 n/n$. Il convient toutefois de retenir que α joue le rôle de potentiomètre en terme de rapport signal sur bruit (voir l'étude numérique de [12]) et doit être choisi proche de 0 pour des valeurs de ϵ elevées et inversement plus grand lorsque le rapport signal sur bruit est élevé.

3.3 Modèle de bruit blanc, loi des déformations inconnue

3.3.1 Problématique

Les paragraphes de la section 3.2 ont apporté des réponses concernant la faisabilité d'estimer f lorsque la loi des déformations est connue, dans des modèles de déformation du type (3.6) et (3.10) lorsqu'on controle n le nombre de signaux mais que le niveau de bruit ϵ est un paramètre figé du problème. L'hypothèse de la connaissance de la loi g peut être satisfaisante si l'on considère par exemple des images médicales enregistrées par tomographie et reconstruction de Radon : si on a préalablement calibré des lois de déformations par rotation ou homothétie sur des patients "tests", il est envisageable d'en déduire une estimation des lois de déformations. De même, si on considère des photos prises par un robot, il on peut estimer la loi des rotations de l'objectif du robot autour de sa position théorique par le biais d'un calibrage en amont de l'appareil. Cependant, ces situations ne sont pas toujours satisfaisantes et il est naturel de se poser la question de la faisabilité d'estimer f sans connaître la loi g.

Ceci peut avoir deux motivations : la première serait d'obtenir au travers des observations des informations sur la structure générative des données et ainsi effectuer (par exemple) des analyses descriptives des modes de variations des signaux. La seconde motivation est que s'il est possible de retrouver les paramètres des déformations (non observées) dans les modèles précédents, alors on peut accéder au signal f lui-même. Pour simplifier, supposons observés à nouveau des signaux selon le modèle de bruit blanc translaté aléatoirement

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

Une méthode pour estimer f pourrait consister à retrouver les paramètres de déformation de chaque signal par le biais d'estimateurs $(\hat{\tau}_j)_{j\in\{1\dots n\}}$, puis inverser la déformation estimée $\hat{\tau}_j$ sur chaque signal Y_j avant d'effectuer une simple moyenne. L'estimateur de f prendrait alors la forme de

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

Dans la suite, nous nous intéresserons donc aux deux questions :
- Est-il possible de retrouver les paramètres de déformation ?
- Est-il possible d'estimer f sans la connaissance de g ?

Remarque 3.3.1 *Il s'agit alors ici de bien préciser le contexte puisque la première question a déjà été traitée dans un cas semi-paramétrique où les courbes $(Y_i)_{i\in\{1\dots n\}}$ sont observées sur une grille échantillonnée de plus en plus finement (voir [Gamboa et al., 2007b], [Bigot et al., 2009] ou [Vimond, 2010]). Notons que considérer que l'échantillonage de la courbe est de plus en plus important revient moralement à placer un niveau de bruit ϵ de plus en plus petit.*

3.3.2 Estimation de f par moyenne de Fréchet

Il est possible d'envisager une procédure globale en utilisant la moyenne au sens de Fréchet de variables aléatoires $Z_1, \dots, Z_n$ qui ne sont plus dans un espace vectoriel V mais dans un espace modélisant des formes translatées. Cette remarque est consistante avec le fait que Z, Z' peuvent être considérées comme identiques si on peut construire des transformations d'un groupe H qui envoie exactement Z sur Z'. Dans [Frechet, 1948], la notion de moyenne euclidienne est étendue aux espaces métriques par le biais d'un critère implicite : si on considère une distance d définie sur une variété $\mathcal{M}$, alors la moyenne de Fréchet de n observations $(Z_i)_{i\in\{1\dots n\}}$ dans $\mathcal{M}$ est définie par

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

En revenant à notre cas des courbes translatées aléatoirement, $H = \mathbb{R}$ est le groupe des translations agissant sur les fonctions $f \in L^2([0,1])$ de période 1 par

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

Ainsi, étant données n réalisations $Y_1, \dots, Y_n$ du modèle (3.6), la moyenne de Fréchet sous l'action de H sera définie par

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

En considérant à nouveau les coefficients des observations (notés $\theta_{m,\ell}$ pour la fréquence ℓ de l'observation m), et en tronquant à une fréquence ℓ_0, on obtient une estimation $(\hat{\theta}_k)_{-\ell_0\leqslant k\leqslant \ell_0}$ par

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

La moyenne de Fréchet est alors obtenue par $\hat{f}^F_{n,\ell_0}(x) = \sum_{|\ell|\leqslant\ell_0} \hat{\theta}_\ell e^{-2i\ell\pi x}$. En remarquant simplement que nécessairement $\hat{\theta}_\ell = \frac{1}{n}\sum_{m=1}^n \theta_{m,\ell} e^{2i\ell\pi\hat{\tau}_m}$, on en déduit que

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

En fin de compte, le calcul de la moyenne de Fréchet revient à la minimisation du critère défini par (3.14), et cela peut être effectué en utilisant une descente de gradient.

3.3.3 Estimation des paramètres de translation

Rappelons que le modèle (3.6) s'écrit statistiquement

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

où $z_{\ell,m}$ sont i.i.d. $\mathcal{N}_{\mathbb{C}}(0,1)$ et les $\tau^*_m, m = 1,\dots,n$ sont ici les vrais paramètres de translation qui ont été tirés selon g. Le problème (3.15) est clairement non identifiable puisque pour tout $\tau_0\in\mathbb{R}$, on peut substituer $\theta_\ell e^{i2\pi\ell\tau_0}$ à θ_ℓ et $\tau^*_m - \tau_0$ à τ^*_m sans changer d'observations. On introduit donc les deux conditions d'identifiabilité suivante :

(H_g) : g est à support compact inclus dans $\mathcal{T} = [-\frac{1}{4},\frac{1}{4}]$ et de moyenne nulle.
(H_f) : La fonction f est telle que $c_1(f)\neq 0$.

L'hypothèse (H_g) nous conduit à nous intéresser à des estimations $(\hat{\tau}_1,\dots,\hat{\tau}_n)$ de moyenne nulle. On introduit donc l'ensemble

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

et si la fréquence de coupure ℓ_0 est fixée, on cherche pour tout $\tau = (\tau_1,\dots,\tau_n)\in\overline{\mathcal{T}}_n$ à optimiser $M_n(\tau)$ puisque c'est ce qui est nécessaire pour construire la moyenne de Fréchet $\hat{f}^F_n$. On peut alors démontrer le résultat suivant de reconstruction des paramètres de translations.

Théorème 3.3.1 *Supposons* (H_g) *et* (H_f) *et définissons l'estimateur*

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

alors pour tout $t>0$

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

avec $C(f,\ell_0,\epsilon,n,t,g) = 4\max\left[C_1(f,\ell_0)\left(\sqrt{C_2(\epsilon,n,\ell_0,t)} + C_2(\epsilon,n,\ell_0,t)\right), C_3(t,n,g)\right]$, $C_1(f,\ell_0)$ *est une constante positive ne dépendant que de* f *et de la fréquence de coupure* ℓ_0,

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

et

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

Le théorème 3.3.1 donne une borne supérieure en probabilité pour la précision de l'estimateur $\hat{\tau}$ par rapport aux vrais paramètres de translations $\tau_m^*, m = 2, \ldots, n$. Le minimum de $M_n(\tau)$ étant calculé sur $\overline{\mathcal{T}}_n$, on obtient $\hat{\tau}_1 = -\sum_{m=2}^n \hat{\tau}_m$. Lorsque $n \to +\infty$, $C(f, \ell_0, \epsilon, n, t, g)$ utilisée dans (3.16) converge vers $4C_1(f, \ell_0)\left(\epsilon^2(2\ell_0+1) + \epsilon\sqrt{2\ell_0+1}\right)$. Ainsi, ceci ne peut donner de la consistance puisque (3.16) ne peut aboutir à $\lim_{n\to+\infty} \frac{1}{n}\sum_{m=2}^n (\hat{\tau}_m - \tau_m^*)^2 = 0$ en probabilité. Au contraire, (3.16) semble suggérer l'existence de $C > 0$ telle que $\frac{1}{n}\sum_{m=2}^n (\hat{\tau}_m - \tau_m^*)^2 > C\epsilon^2(2\ell_0+1)$ se produit avec une probabilité strictement positive. Cela montrerait donc que la précision de $\hat{\tau}$ va dépendre du niveau de bruit ϵ^2 et de la fréquence de coupure ℓ_0.

3.3.4 Borne inférieure de 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

$(\tilde{H}_f)$: f est telle que $\|f'\|_2^2 = \sum_{\ell\in\mathbb{Z}} (2\pi\ell)^2 |c_\ell(f)|^2 < +\infty$.

$(\tilde{H}_g)$: La densité g est à support compact $\mathcal{T}$ avec $\lim_{\tau\to\inf\mathcal{T}} g(\tau) = \lim_{\tau\to\sup\mathcal{T}} g(\tau) = 0$.

On a alors le théorème

Théorème 3.3.2 *Soit $X = (\theta_{m,\ell})_{\ell\in\mathbb{Z}, m=1,\ldots,n}$ l'ensemble des coefficients de Fourier observés dans $\mathbb{X} = \ell^2(\mathbb{Z})^{\otimes n}$ et soit $\hat{\tau}^n = \hat{\tau}^n(X) \in \mathbb{X}$ une fonction mesurable de X quelconque. Alors, si $(\tilde{H}_f)$ et $(\tilde{H}_g)$ sont vérifiées, on a*

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

où $I(g)$ est l'information de Fisher :

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

Le résultat donné par le théorème 3.3.2 s'appuie sur l'inégalité de van Trees qui est une inégalité de Cramer-Rao bayésienne. Lorsque $n \to +\infty$, $\mathbb{E}\left(\frac{1}{n}\sum_{m=1}^n (\hat{\tau}_m^n - \tau_m^*)^2\right)$ ne peut converger vers 0, ce qui explique le résultat donné par le théorème 3.3.1. Il est possible d'affaiblir l'hypothèse $f \in \mathcal{C}^1([0,1])$ en considérant les estimateurs $\hat{\tau}^{n,\ell_0}$ construits à partir des $\theta_{m,\ell}$ pour $m = 1, \ldots, n$ et $|\ell| \leqslant \ell_0$ dans le modèle (3.15). Dans ce cas, le théorème donne la borne inférieure suivante.

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

3.3.5 Reconnaissance de forme moyenne par modèles déformables

L'approche de moyenne de Fréchet peut être étendue au cas des images qui subissent des déformations plus générales que des éléments d'un groupe de Lie. Nous proposons dans [11] un modèle aléatoire de déformation élastique décrit par l'équation générant les difféomorphismes (3.3). En considérant des images de niveau de gris définies sur $\Omega \subset \mathbb{R}^2$, une image I est donc une application $I : \Omega \longmapsto \mathbb{R}$. Notre objectif est alors d'interpréter le modèle de moyenne de Fréchet comme un M-estimateur (voir par exemple [Van der Waart, 1998] pour de nombreux détails sur ces estimateurs) dans ce contexte précis.

Tout d'abord, nous proposons un modèle qui génère des difféomorphismes aléatoires paramétriques en utilisant l'approche développée par [Trouvé and Younes, 2005] pour des champs de vecteurs homogènes en temps dans l'équation (3.3). Sans perte de généralité, on choisit $\Omega = [0,1]^2$ et nous imposons une structure paramétrique pour une fonction $v : [0,1]^2 \mapsto \mathbb{R}^2$ vérifiant $v_{\partial[0,1]^2} = 0$. Ainsi, si $(e_1, \dots e_K)$ désigne une famille finie de fonctions de base de $[0;1]^2$ dans $\mathbb{R}^2$ s'annulant aux bords de $[0,1]^2$, on obtient un champ de vecteur aléatoire v_a en générant aléatoirement 2K coefficients $(a_1^1, \dots a_K^1) \times (a_1^2, \dots a_K^2)$ pour lesquels

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

Étant donné v_a, on obtient alors un difféomorphisme aléatoire en utilisant simplement la solution au temps 1 de (3.3) notée $\Phi^1_{v_a}$. La construction d'un modèle d'image déformée aléatoirement peut alors s'écrire simplement. Étant donnée une loi P_A à support compact dans $[-A, A]$, où $A > 0$ et un entier K quelconque, on définit une image déformée aléatoirement par

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

où ϵ est un bruit additif indépendant des coefficients a de loi $P_A^{\otimes 2K}$. On suppose alors observées n réalisations indépendantes de (3.17) notées I_{ϵ_i,a_i}. On notera V_A l'ensemble des champs de vecteurs pouvant être générés avec des coefficients variants dans $[-A, A]^{2K}$ dans la définition de v_a. Pour une image Z définie sur $[0,1]^2$, on note la fonction de contraste utilisant une discrétisation $\mathcal{P}$ en pixel de $[0,1]^2$:

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

Par conséquent, f mesure le coût $\ell^2(\mathcal{P})$ d'alignement de Z sur $I_{\epsilon,a}$ en utilisant un difféomorphisme de V_A. La fonction de contraste moyenne est alors

$$F(Z) = \int f(a, \epsilon, Z) d\mathbb{P}(a, \epsilon),$$

et la moyenne de Fréchet intrinsèque de la variable aléatoire $I_{\epsilon,a}$ est définie par $Q^* = \arg\min_{Z \in \mathcal{Z}} F(Z)$. Si $\mathbb{P}_n$ désigne la mesure empirique des observations, il est possible de définir le contraste empirique par

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

On définit donc la moyenne de Fréchet des observations comme étant $\hat{Q}_n = \arg\min_{Z \in \mathcal{Z}} F_n(Z)$ et cet estimateur est calculable contrairement à Q^* puisque la loi des déformations est ici inconnue.

Bien entendu, la minimisation de (3.18) peut aboutir à des estimateurs très différents de I^* et le contraste F_n doit être régularisé en pratique. Dans [11], nous adjoignons à F_n un terme de pénalisation permettant de contrôler la régularité de Z ainsi que la quantité de déformation possible pour procéder aux recalages et démontrons alors des propriétés de convergence presque sûre des estimateurs de moyenne de Fréchet $\hat{Q}_n$ vers Q^*.

Il est important de noter que nous ne savons pas si I^* appartient à Q^*, et de ce fait la réponse apportée par notre méthode est encore très partielle dans cette situation.

3.4 Résultats numériques

3.4.1 Modèle de courbes translatées aléatoirement

Nous donnons quelques résultats numériques sur le problème de l'estimation de la forme moyenne f lorsque les données sont issues du modèle de translation aléatoire. Quatre fonctions f sont étudiées (voir Figures 3.5(a)-3.8(a)) et nous utilisons $n = 200$ courbes bruitées et translatées aléatoirement avec une loi de Laplace de densité $g(x) = \frac{1}{\sqrt{2}\sigma} \exp\left(-\sqrt{2}\frac{|x|}{\sigma}\right)$ avec $\sigma = 0.1$. Un échantillon de 10 courbes est donné dans les Figures 3.5(b)- 3.8(b) pour chaque fonction moyenne à estimer. Enfin, nous donnons l'estimation de f par la simple moyenne arithmétique dans les Figures 3.5(c)- 3.8(c). On constate immédiatement la convolution par g de cette estimation qui est bien loin de donner quelque chose de satisfaisant.

Les coefficients de Fourier de g sont donnés par $\gamma_\ell = \frac{1}{1+2\sigma^2\pi^2\ell^2}$, ce qui correspond à un ordre de problème inverse $\nu = 2$. Nous donnons dans les Figures 3.5(d)(e) -3.8(d)(e) l'estimation par problème inverse $\widehat{f}_n^H$ décrit dans le paragraphe 3.2.1 (et on se reportera à [9] pour plus de détails). Enfin, l'estimation sans connaître la loi des déformations est donnée dans les Figures 3.5(f) -3.8(f) par la méthode de Fréchet décrite dans le paragraphe 3.3.2 (estimation $\widehat{f}_n^F$). On constatera l'efficacité des deux dernières méthodes, notamment même lorsqu'on ne connait pas la loi des déformations et qu'il faut alors estimer les coefficients de Fourier de g.

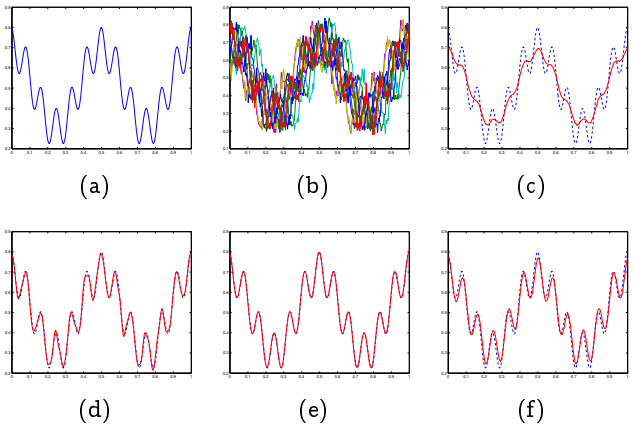


FIGURE 3.5 – Fonction "Wave". (a) Vraie fonction f, (b) Échantillon de 10 courbes parmi $n = 200$, (c) Moyenne empirique, Déconvolution (d) $\widehat{f}_{n,1}^H$ et (e) $\widehat{f}_{n,2}^H$, (f) Moyenne de Fréchet

3.4.2 Moyenne de Fréchet d'images

Nous illustrons ici les travaux décrits dans le paragraphe 3.3.5. Pour l'image classique de Lena la Figure 3.9 donne deux réalisations du modèle de déformation par flot de difféomorphisme lorsque les coefficients a_k sont tirés uniformément sur $[-A, A]$. La quantité de déformation autorisée dans ce modèle est donc proportionnelle à A et les fonctions B-spline utilisées pour paramétrer les champs de vecteur permettent de localiser les effets de déformation.

Enfin, nous comparons notre méthode de recalage par moyenne de Fréchet et action de flots de difféomorphismes avec la simple moyenne empirique sur deux célèbres base de données : la base Mnist de chiffres manuscrits (voir Figure 3.10 pour l'effet de l'algorithme décrit dans [11] sur le chiffre "2") et la base Olivetti (voir Figure 3.11 pour l'utilisation sur des recalages de visages).

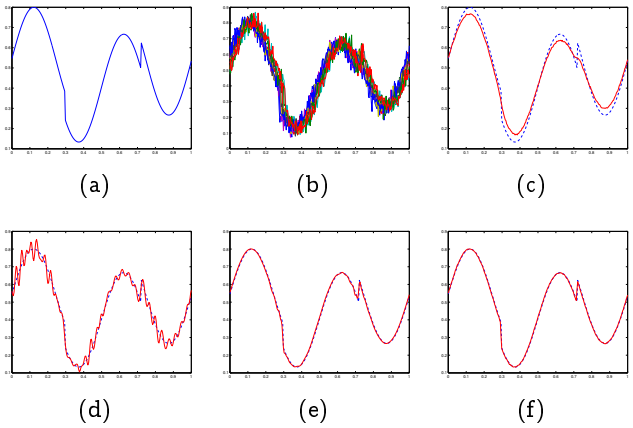


FIGURE 3.6 – Fonction HeaviSine. (a) Vraie fonction f, (b) Échantillon de 10 courbes parmi $n = 200$, (c) Moyenne empirique, Déconvolution (d) $\widehat{f}^{H}_{n,1}$ et (e) $\widehat{f}^{H}_{n,2}$, (f) Moyenne de Fréchet

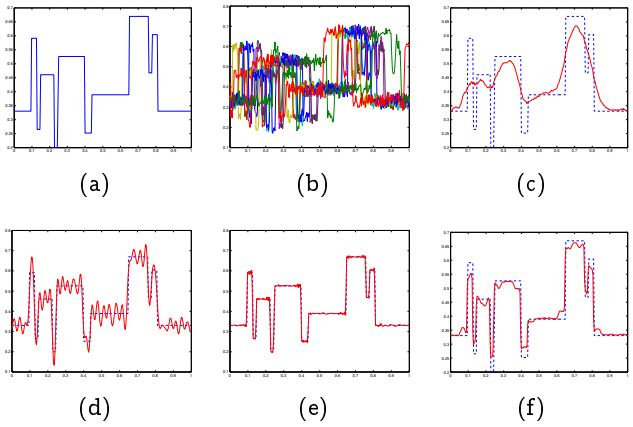


FIGURE 3.7 – Fonction Blocks. (a) Vraie fonction f, (b) Échantillon de 10 courbes parmi $n = 200$, (c) Moyenne empirique, Déconvolution (d) $\widehat{f}^{H}_{n,1}$ et (e) $\widehat{f}^{H}_{n,2}$, (f) Moyenne de Fréchet

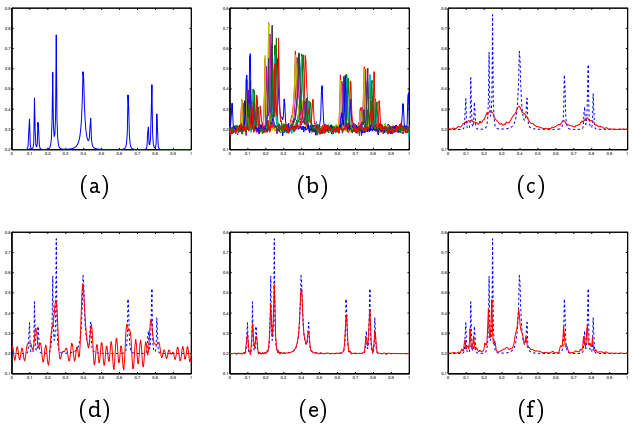


FIGURE 3.8 – Fonction Bumps. (a) Vraie fonction f, (b) Échantillon de 10 courbes parmi $n = 200$, (c) Moyenne empirique, Déconvolution (d) $\widehat{f}^{H}_{n,1}$ et (e) $\widehat{f}^{H}_{n,2}$, (f) Moyenne de Fréchet

Notons enfin la possibilité d'aménager cet algorithme pour effectuer un algorithme de clustering d'images. Nous renvoyons à [11] pour plus de détails sur ces expériences.

FIGURE 3.9 – Déformation aléatoire de Lena avec $A = 0.1$ et $A = 0.5$.

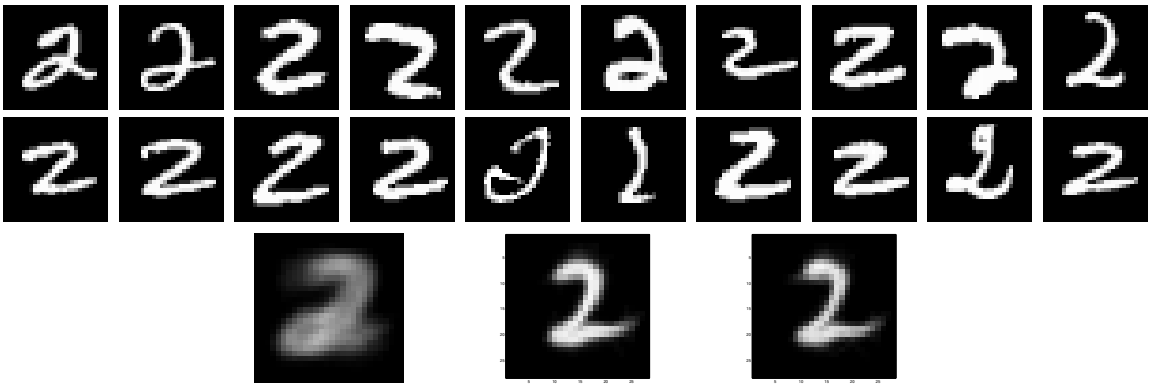


FIGURE 3.10 – Moyenne empirique (en bas à gauche), Moyennes $Z_\star^{(1)}$ à la première itération de l'algorithme et à l'itération 3 $Z^{(3)}$ de l'algorithme décrit dans [11].

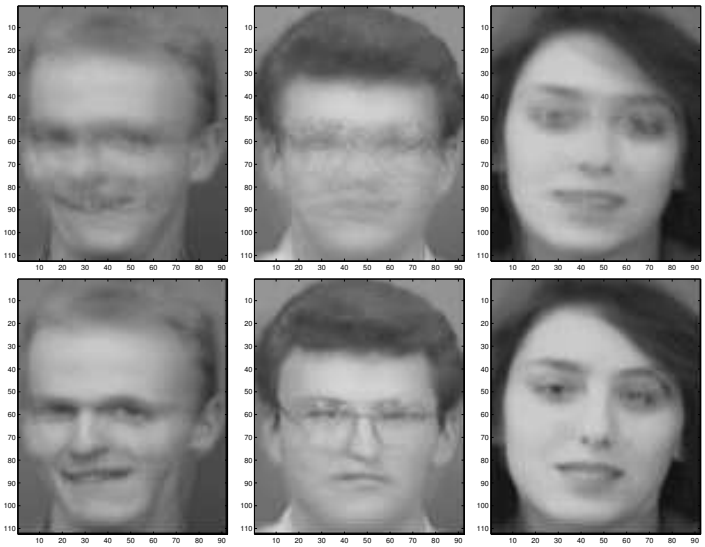


FIGURE 3.11 – Exemple de moyenne de Fréchet obtenue sur 3 visages de la base Olivetti. Première ligne : moyenne empirique, seconde ligne : moyenne de Fréchet.

3.5 Élargissements

3.5.1 Régression sous contrainte de forme

L'estimateur de régression monotone développé au travers de l'approche par difféormorphismes issus de champs de vecteurs est numériquement extrêmement performant. Il conviendrait sans

doute de mieux le diffuser en développant une bibliothèque logicielle.

Par ailleurs, la question de savoir si la paramétrisation en termes de réalisation au temps 1 de flots de difféomorphismes peut s'étendre au cas où la fonction à estimer est convexe. Ce problème de régression sous contrainte de convexité possède en effet des applications en finance [Ait-Sahalia and Duarte, 2003] pour fixer les prix de stock-options, en économie [Allon et al., 2007] où la demande, la production sont souvent concaves, ainsi qu'également en biologie [Ratkowsky, 1983], en analyse de surfaces de réponse pour des problèmes d'optimisation [Hoffmann et al., 2006]. Si la régression doit être effectuée dans $\mathbb{R}^d$, il s'agirait par exemple de considérer une initialisation $\phi_0(x) = |x|^2$ et une évolution $\frac{d\phi_t}{dt} = v_t(\phi_t)$ où v_t préserve la convexité et telle que ϕ_1 matche la fonction f à estimer.

Enfin, d'un point de vue plus théorique, la question d'affiner la convergence du théorème 3.1.2 est toujours en suspens. En effet, ous prouvons dans [10] un résultat intermédiaire donnant simplement une convergence en probabilité de $(v_t^{n,\lambda})$ vers v_t au travers de techniques standard de M-estimation. Il est envisageable de s'appuyer sur des résultats plus fins de M-estimation pour des éléments de dimension infinie en s'appuyant sur des résultats de type Donsker (voir par exemple [van der Vaart and Wellner, 1996]). L'intérêt d'une telle approche consisterait alors à l'élaboration de nouveau tests statistiques de monotonie et convexité.

3.5.2 Approche Bayésienne dans le modèle déformé avec opérateur inconnu

Afin de simplifier la problématique, je limite le paragraphe volontairement à l'étude des courbes translatées aléatoirement donné en (3.6) :

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

où la loi g est *inconnue*. Nous avons vu dans le théorème 3.3.2 que sans un moyen de faire baisser le niveau de bruit ϵ vers 0, il n'est pas possible de retrouver les paramètres de déformation, ce qui de fait rend quasiment impossible la consistance de la moyenne de Fréchet pour l'estimation de f à la vue des estimations données par (3.13) et (3.14). Par contre, il ne semble pas impossible d'estimer f sans pour autant estimer les paramètres de déformation, ceci en considérant une méthode bayésienne : c'est par exemple l'approche utilisée dans [Allassonière et al., 2007] et [Allassonière et al., 2009] sans pour autant que la consistance soit démontrée dans ce contexte, même dans un problème d'estimation paramétrique.

En revenant aux origines des travaux de [Ibragimov and Has′minskiĭ, 1981, Le Cam, 1973] portants sur la consistance bayésienne, il existe des conditions suffisantes de consistance des méthodes bayésiennes dans le cas où le modèle ne comporte pas de variables cachées. Essentiellement, étant données des observations $(X_1, \dots X_n)$ de loi P_{θ_0} où le paramètre $\theta_0 \in \Theta \subset \mathbb{R}^d$ est inconnu, l'estimateur bayésien de θ_0 dépend d'un a priori noté q qui est une distribution de probabilités sur Θ. Si les lois $(P_\theta)_{\theta\in\Theta}$ sont à densité notées $(p_\theta)_{\theta\in\Theta}$ par rapport à la mesure de Lebesgue sur $\mathcal{X}$ l'espace des réalisations, l'estimateur bayésien est donné par

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

où L est une fonction de perte s'annulant en $0_{\mathbb{R}^d}$, typiquement $L(x) = |x|^p$ pour $p > 0$. Le fameux théorème 5.2 du chapitre 1 de [Ibragimov and Has′minskiĭ, 1981] prouve que l'on peut espérer la consistance de l'estimation si le rapport de vraisemblance des n observations varie assez régulièrement entre deux valeurs du paramètre θ et devient petit loin du vrai paramètre θ_0 du moment que q est continue sur Θ et strictement positive. C'est par ailleurs précisément la régularité de cette variation du rapport de vraisemblance qui donne la vitesse de recontruction

de $\hat{\theta}_n^B$ autour de θ_0 et cette régularité est liée à la séparabilité des lois $(P_\theta)_{\theta\in\Theta}$ en distance de Hellinger [2]. À noter que la consistance de l'estimateur bayésien s'interprète comme un cas particulier de méthode de Laplace pour les équivalent d'intégrale où la distribution a posteriori de θ se concentre autour d'une loi gaussienne autour de θ_0 avec une variance tendant vers 0 (voir par exemple le théorème de Bernstein-von Mises dans (décrit dans [Le Cam, 1973] ou [Van der Waart, 1998]).

D'un point de vue de statistiques mathématiques non paramétriques, la situation se complique naturellement. Certains développements [Ghosal et al., 2008, Ghosal, 2000, Rousseau, 2010] proposent des extensions permettant des résultats dans un contexte ou la dimension de θ devient virtuellement infinie. La technique statistique s'appuie sur des minorations de distance de Hellinger par des distances se rapport à l'estimation sur Θ, ainsi qu'à des arguments de recouvrement. Parmi la littérature grandissante sur le sujet, on se rapportera aux travaux de [Ghosal et al., 2008, Ghosal, 2000, Rousseau, 2010]. L'idée principale pour obtenir des estimations non paramétriques adaptatives est de construire des *sieves* convenablement.

La difficulté supplémentaire dans notre contexte (3.6) (et qui était déjà présente pour l'estimation fréquentiste) est que les observations dépendent d'un paramètre caché qui est la translation aléatoire tirée selon g. En imaginant qu'un seul coefficient de Fourier θ_0 de f doit être estimé, les coefficients de Fourier des observations sont

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

Si q désigne un a priori sur Θ et r un a priori sur $\mathcal{L}^1(\mathcal{S}^1)$ (espace des densités sur la sphère de dimension 1 paramétrée par $e^{2i\pi\tau}$), il est à nouveau possible de formuler une estimation bayésienne de (θ_0, g) :

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

où $I_{u,v}(X_1, \dots, X_n)$ désigne le rapport de vraisemblance associé aux a priori q, r. Un bon angle de vue pour contrôler ce modèle de mélange « infini »pourrait s'inspirer des travaux dans[Rousseau, 2010].

Un premier point à vérifier pour obtenir un résultat de consistance serait de vérifier l'identifiabilité dans le modèle Bayésien, à la fois pour f et g. Un tel résultat devrait découler du lien qui existe entre la variation totale entre les lois $d_{VT}(\mathbb{P}_{f,g}, \mathbb{P}_{\tilde{f},\tilde{g}})$ et la transformée de Laplace $\mathcal{L}(g-\tilde{g})$ et les coefficients de Fourier de f et $\tilde{f}$. Ensuite, il est nécessaire pour l'estimation non paramétrique de contrôler les nombres de recouvrement des lois $\mathbb{P}_{f,g}$ pour la distance de Hellinger (ou bien pour la distance de Kullback Leibler ou en variation totale). Ceci constitue le coeur de la difficulté puisqu'il ne peut exister d'inégalités entre ces distances de lois et des distances sur les espaces intrinsèques dans lesquels vivent $f, \tilde{f}, g, \tilde{g}$. Il est donc impératif de calculer des bornes supérieures de nombres de recouvrement pour les mélanges infinis de Gaussiennes puisque c'est ici ce qui génère nos données.

3.5.3 Modèle de bruit Poissonien

Dans [17], nous étendons l'étude asymptotique des courbes translatées lorque nous sommes dans un contexte de l'estimation d'une intensité λ d'un processus de Poisson. Les observations sont cette fois données par des processus de comptage définis sur $[0,1]$ $N^1, \dots N^n$ d'intensités respectives $\lambda(.-\tau_1), \dots, \lambda(.-\tau_n)$ où les translations aléatoires non observées suivent une loi g connue et sont tirées indépendament. Un tel modèle peut être utilisé dans le contexte des données

2. En réalité, la distance de Kullback semble même suffisante, ce qui est une hypothèse légèrement plus faible.

Chip-Seq comme indiqué dans l'introduction. Nous démontrons dans [17] que l'estimation de λ ne peut être meilleure que dans la situation de bruit blanc gaussien. Plus précisément, nous obtenons le théorème suivant.

Théorème 3.5.1 *Supposons que la densité g satisfait une hypothèse de décroissance de ses coefficients de Fourier (3.8). Si*

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

Soit $A > 0$ et supposons que la régularité de λ notée s est telle que $s > 2\nu + 1$. Alors il existe $C_0 > 0$ (indépendante de n) telle que pour n assez grand

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

où le minimum est pris sur tous les estimateurs $\hat{\lambda}_n \in \Lambda_0$ (i.e. fonctions mesurables des processus N^i, $i = 1, \dots, n$ à valeurs dans Λ_0).

La technique de démonstration de cette borne inférieure repose encore une fois sur l'analyse fine du rapport de vraisemblance entre deux hypothèses λ et $\lambda + h$ couplée à une approche utilisant le Lemma d'Assouad. Le rapport de vraisemblance s'écrit par le biais d'un changement de mesure pour les procesus de Poisson et en repassant à nouveau par l'intermédiaire d'une hypothèse « nulle », c'est à dire qui est invariante par translation. Cette fois-ci, l'hypothèse nulle à utiliser est celle d'une intensité de processus de comptage constante égale à ρ sur $[0,1]$. Ainsi, si λ_1, λ_2 sont deux intensités telles que $\lambda_1 \geqslant \rho > 0$ et $\lambda_2 \geqslant \rho > 0$, et si on observe un processus de comptage N, la vraisemblance s'écrit

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

où $\mu_1 = \lambda_1 - \rho$ et $\mu_2 = \lambda_2 - \rho$.

Puis, nous proposons à nouveau un estimateur adaptatif $\hat{\lambda}^h_n$ basé sur une méthode de seuillage dur qui atteint asymptotiquement la borne inférieure à un facteur logarithmique près $\mathcal{O}\left(\left(\frac{\log n}{n}\right)^{\frac{2s}{2s+2\nu+1}}\right)$. Cet estimateur est précisément décrit dans [17].

Notons enfin qu'il semble extrêmement pertinent d'exploiter cette étude théorique d'un point de vue pratique puisque l'utilisation de l'estimation précédente permettrait d'obtenir des procédures de recalage automatiques pour les données de comptage haut débit Chip-Seq et que ce problème semble devenu incontournable dans la manipulation de ce genre de données.

3.5.4 Problèmes de tests

Enfin, il semblerait raisonnable d'exploiter la structure de rapport de vraisemblance dans les modèles poissoniens et de bruit blanc pour en déduire des procédures de test dans le problème qui étudierait si différentes observations sont issues ou non d'un modèle de bruit blanc ou d'un modèle de processus de comptage translaté aléatoirement. Une approche pour démarrer l'étude de cette question est décrite dans un cadre plus simple dans les travaux [Fromont et al., 2011] et pourrait consister en un point de démarrage prometteur pour ce problème.

Chapitre 4

Algorithmes d'optimisation non réversible

Je détaille dans ce chapitre mes travaux inspirés du système dynamique décrit par l'équation différentielle

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

où U est un potentiel défini sur $\mathbb{R}^d$ coercif en l'infini : $\lim_{|x|\mapsto+\infty} U(x) = +\infty$. Rappelons que l'étude de ce système différentiel provient d'un aménagement numérique détaillé dans le paragraphe 1.4 s'intéressant à la minimisation de U. On supposera sans perte de généralités que $\min_{\mathbb{R}^d} U > 0$, que U est au moins $\mathcal{C}^2(\mathbb{R}^d)$ et convexe à l'infini [1], au sens où

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

Par ailleurs, on supposera que $U(0) = \min U$.

4.1 Modèle de descente de gradient à mémoire

4.1.1 Lien avec un problème physique

Étant données h et k deux fonctions régulières croissantes et positives, on considère l'équation différentielle sur $\mathbb{R}^d$ donnée par

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

Un cas particulièrement naturel sera le cas où $k \sim \int h$ pour les grandes valeurs de t. En effectuant un changement de temps $t \mapsto \tau(t)$, il est possible de transformer (4.1) en une équation du second ordre.

Proposition 4.1.1 *Si τ désigne la solution de $\dot{\tau} = \sqrt{k(\tau)/h(\tau)}$ et x est solution de (4.1), alors $z = x \circ \tau$ est solution de*

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

où a est une fonction d'amortissement donnée par $\gamma(s) = \left(\frac{\dot{k}h+k\dot{h}}{2h^{3/2}k^{1/2}}\right)\circ\tau(s)$.

1. Nous interpréterons en fait cette condition comme une condition de rappel du système (4.1)

La fonction γ quantifie donc un frottement dans une dynamique d'une boule roulant sur le graphe de la fonction U, et soumise à l'apesanteur. Certains cas de fonctions mémoires h et k permettent alors de retrouver le système *Heavy Ball with Friction* et (4.1) en est donc une généralisation.

À partir de cette interprétation, on peut notamment imaginer que le système paramétré par (4.2) se stabilise sur un minimum de U et que l'inertie accumulée par la trajectoire permette à la descente de gradient moyennée de sauter certains maxima locaux, contrairement à une descente de gradient classique.

Tout d'abord, on montre que les trajectoires solutions de (4.2) sont stables par le biais d'une fonction de Lyapunov classique $\mathcal{E}$ décrite ci-dessous dès que U satisfait une hypothèse de convexité à l'infini de U.

4.1.2 Comportement du système dynamique (4.2), cas convexe

On se place dans un cadre un peu plus général que celui de la convexité décrit par l'hypothèse [2] :

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

Rôle de l'amortissement γ L'effet de l'amortissement γ est important et doit être compris ainsi : si l'amortissement est trop rapidement décroissant vers 0, la trajectoire (4.2) présente une infinité d'oscillations non négligeables, puisque le système se rapproche alors d'une équation du type $\ddot{x} + \omega^2 \nabla U x = 0$. Ce résultat est décrit dans la proposition suivante :

Proposition 4.1.2 *Si on note $\mathcal{E}(t) = U(x(t)) + \frac{\dot{x(t)}^2}{2}$, alors $\dot{\mathcal{E}}(t) = -\gamma(t)|\dot{x}(t)|^2$ et les solutions de (4.2) sont définies sur $\mathbb{R}_+$ et restent bornées. Par ailleurs,*

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

Ainsi, même si U est convexe, si $\gamma \in \mathbb{L}^1(\mathbb{R}_+)$ la trajectoire ne peut converger.

La proposition précédente nous incite donc à considérer un amortissement $\gamma \notin \mathbb{L}^1(\mathbb{R}_+)$. On peut alors donner une condition suffisante de convergence de $(U(x_t))_{t \geqslant 0}$.

Proposition 4.1.3 *Supposons que U vérifie la condition (H_U^1).*

i) *Si γ est C^1 et décroissante, alors*

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

ii) *Si de plus $\int_0^{+\infty} \gamma(s)ds = +\infty$ (cas de décroissance lente), alors $\lim \mathcal{E}(t) = \min U$.*

iii) *S'il existe $m > 0$ tel que $\gamma(t) \geqslant m/t$ pour t assez grand, alors*

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

iv) *Si enfin $\arg\min(U) = \{0\}$, la trajectoire converge.*

Notons que cette proposition ne donne pas de propriété de convergence de la trajectoire elle-même dans la situation où $\arg\min U$ contiendrait un ouvert de $\mathbb{R}^d$. Dans cette situation, il est possible de donner une hypothèse quasiment minimale.

2. Cette condition est satisfaite dès que U est convexe avec $\theta = 1$ par exemple.

Théorème 4.1.1 *Supposons que* $d = 1$ *et que* U *satisfait* (H_U^1) *avec* $[\alpha, \beta] \subset \arg\min U$. *Si* γ *satisfait*

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

alors la trajectoire solution de (4.2) converge. Si par contre γ *satisfait*

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

la trajectoire ne converge pas, sauf si elle est initialisée à vitesse nulle dans $\arg\min(U)$.

Il est également possible de démontrer un résultat de non convergence similaire en dimension supérieure, ces détails sont omis car les prérequis sont un petit peu techniques et mais sont disponibles dans [13]. L'hypothèse clef de convergence ou non est décrite par la condition $\int_0^{+\infty} e^{-\int_0^s \gamma(u)du} ds = \infty$. Notons enfin que la preuve fait intervenir une fonction de Lyapunov permettant de contrôler vitesse et position de la trajectoire, ce qui n'est pas le cas de $\mathcal{E}$ comme l'indique la proposition 4.1.2). Cette fonction fait intervenir en plus des termes de $\mathcal{E}$ un terme produit entre vitesse et position et est relativement classique lorsqu'on étudie des systèmes dissipatif (voir [Haraux, 1991] par exemple). Cette adjonction de termes produits a depuis été très largement utilisée dans les contextes d'hypocoercivité en équations aux dérivées partielles ou en probabilités.

4.1.3 Comportement du système dynamique (4.2), cas non convexe

Le cadre non convexe est restreint à une situation générique décrite par l'hypothèse suivante $(\tilde{H}_U)$: U possède un nombre fini m de points critiques tels que $U(x_1) < U(x_2) \cdots < U(x_m)$

Sous cette hypothèse, et en se plaçant dans le cas où $\gamma \notin \mathbb{L}^1(\mathbb{R}_+)$, on peut démontrer un résultat plus faible qu'une convergence trajectorielle dans le cas général de la dimension supérieure ou égale à 1.

Théorème 4.1.2 *Supposons que* $(\tilde{H}_U)$ *soit vraie, alors il existe un unique* x_i *tel que*

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

On peut récrire l'ergodicité en

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

Pour le cas de la dimension 1, il est possible d'obtenir un résultat de convergence plus fort dont la preuve est vraiment spécifique à la nature de la dimension 1 et consiste à considérer les changements de signe de $\dot{x}(t)$.

Théorème 4.1.3 *Supposons que tous les points critiques de* U *vérifient* $U''(x_i) \neq 0$ *et que l'amortissement* γ *est minoré par* $\gamma(t) \geqslant \frac{c}{1+t}$ *où* $c > 0$. *Alors*

i) *Quelle que soit l'initialisation, les solutions de (4.2) vérifient* $\lim_{t\to\infty} x(t) = x^\star$ *existe et appartient à* $\{x_1, \dots x_m\}$.

ii) *Si* $\mathcal{T}$ *est l'ensemble des temps de changement de signe de* $\dot{x}$, *alors*

$$|\mathcal{T}| = +\infty \Longleftrightarrow x^\star \textit{ est un minimum (local) de } U.$$

iii) *L'ensemble des points d'initialisation tel que* $x^\star$ *est un minimum local est un ouvert dense.*

Ainsi, il n'a pas été possible de démontrer des résultats plus forts comme la convergence vers le minimum global de U. Nous avons donc naturellement été poussés à étudier une évolution perturbée de (4.1).

4.2 Diffusion renforcée par sa mémoire

4.2.1 Modèle de diffusion moyennée

Nous étudions par la suite l'extension naturelle de (4.1) lorsque l'évolution est perturbée par un mouvement brownien non dégénéré. En considérant toujours h et k deux fonctions croissantes et positives, si σ désigne une matrice de covariance inversible de taille d et $(B_t)_{t\geqslant 0}$ un mouvement brownien standard de dimension d, l'évolution est décrite par l'équation différentielle stochastique

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

Si $(Y_t)_{t\geqslant 0}$ est le processus définissant la vitesse instantanée,

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

on obtient alors $dY_t = (h/k)(t)(\nabla U(X_t) - Y_t)dt$. Ainsi, (4.3) est un système différentiel cinétique $2d$-dimensionnel non homogène et Markovien décrit par les équations couplées :

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

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

Nous nous sommes volontairement limité au cas où $h = \dot{k}$ même s'il est possible d'étendre nos résultats à des mémoires un peu plus générales. Il est rapide de vérifier que $(X_t, Y_t, t)_{t\geqslant 0}$ devient un processus homogène de générateur $\mathcal{A}$ agissant sur $f \in \mathcal{C}^2_K(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}_+)$ par :

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

Par la suite, on supposera que U est un potentiel satisfaisant l'hypothèse (H_U) suivante :

Hypothèse 6 ($\mathbf{H_U}$) $\lim\limits_{|x|\to+\infty} U(x) = +\infty \quad \liminf\limits_{|x|\to+\infty}\langle x, \nabla U(x)\rangle > 0, \quad Tr\left[\sigma^* D^2 U \sigma\right] \leqslant CU.$

Cette hypothèse est satisfaite pour une grande classe de potentiels U : par exemple $U(x) \sim_\infty C_1|x|^p$ avec $D^2U(x) \sim_\infty C_2|x|^{p-2}$ vérifie $(\mathbf{H_U})$ dès que $\|\sigma(x)\| = O(|x|)$. C'est également vrai pour des croissances de U plus faibles : $U(x) \sim_\infty C_1 \ln|x|$ et $D^2U(x) \sim_\infty C_2|x|^{-2}$ avec $\|\sigma(x)\| = O(1+|x|)$ vérifie également $(\mathbf{H_U})$.

Proposition 4.2.1 *Sous l'hypothèse $(\mathbf{H_U})$, il existe une unique solution (au sens fort) de (4.4). De plus, si (X_0, Y_0) vérifie $\mathbb{E}[U(X_0) + |Y_0|^2] < +\infty$, alors pour tout $T > 0$*

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

L'obtention d'une telle propriété repose principalement sur un argument de non explosion en temps fini des trajectoires et est basé sur un contrôle assez grossier obtenu par lemme de Gronwall appliqué au travers de la fonction de Lyapunov 'standard' suivante :

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

La suite de notre étude sur ce thème concerne les propriétés de régularité du semi-groupe associé à $(X_t, Y_t, t)_{t\geqslant 0}$, ainsi que la convergence vers un régime d'équilibre éventuel. Par convention, on notera toujours $z_0 = (x_0, y_0) \in \mathbb{R}^d \times \mathbb{R}^d$ le point d'initialisation (aléatoire ou non) de la diffusion.

4.2.2 Hypo-ellipticité

Le processus (4.4) étant totalement dégénéré sur la coordonnée Y, l'obtention de propriétés d'existence et régularité de la densité $P_t(z_0,.)$ n'est pas immédiate. Nous démontrons deux résultats importants en vue d'établir l'irréductibilité du processus Markovien [3].

Existence et régularité de densité par rapport à la mesure de Lebesgue Le premier de ces résultats concerne l'existence de densité par rapport à la mesure de Lebesgue et utilise les ensembles $\mathcal{E}_U$ définis par

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

On suppose alors que :

Hypothèse 7 ($\mathbf{H_{Hypo}}$) *σ et U sont $\mathcal{C}^\infty$ et il existe $\varepsilon_0 > 0$ tel que $\sigma\sigma^* \geqslant \varepsilon_0 \mathrm{Id}$, (uniforme ellipticité de σ sur $\mathbb{R}^d$). Par ailleurs, la sous-variété $\mathcal{M}_U$ est telle que $\dim(\mathcal{M}_U) \leqslant d-1$.*

On définit les champs de vecteur qui interviennent dans (4.4) en dissociant l'opérateur d'ordre 2 et le champ de vecteur de drift. Ainsi

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

où

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

tandis que :

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

Proposition 4.2.2 *Sous l'hypothèse ($\mathbf{H_{Hypo}}$), alors pour tout $z_0 \in \mathbb{R}^d \times \mathbb{R}^d$ et tout $t > 0$, $P_t(z_0,.)$ est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^d \times \mathbb{R}^d$. De plus, pour tout $t > 0$ et $z_0 \in \mathbb{R}^d \times \mathbb{R}^d$, $z \mapsto p_t(z_0,z)$ est $\mathcal{C}^\infty$ sur $\mathbb{R}^d \times \mathbb{R}^d$ où $p_t(z_0,.)$ est alors la densité de probabilité associée à $P_t(z_0,.)$.*

Cette proposition exploite le fait que sous la condition ($\mathbf{H_{Hypo}}$), l'algèbre de Lie engendrée par $\partial_t + (L_D - L_\sigma), \sigma_1, \dots, \sigma_d$ est de dimension $2d+1$, ce qui rend alors possible l'utilisation du théorème d'Hörmander. Il faut noter enfin que la proposition précédente ne dit rien sur la

3. L'irréductibilité est principalement importante dans le cas où le processus est homogène en temps afin d'obtenir l'unicité de la mesure stationnaire du processus.

régularité de l'application $(z_0, z) \mapsto p_t(z_0, z)$, cette application semble être également continue dès lors qu'on fasse en plus une hypothèse supplémentaire de croissance sur les champs de vecteur, cette question pourrait être abordée par le biais de calcul de Malliavin ou d'inégalités de Harnack (voir par exemple [Hairer, 2011] ou [Pascucci and Polidoro, 2006]) mais n'étant spécialiste ni de l'un, ni de l'autre, ce point est resté en suspens et a été contourné dans la suite de l'étude.

Minoration de la densité $p_t(z_0, .)$ La question de la positivité de $p_t(z_0, .)$ précédemment introduite est finalement assez différente d'une simple utilisation de la condition de Hörmander pour l'existence d'une densité régulière. En effet, minorer $p_t(z_0, z)$ signifie finalement trouver une quantité suffisante de trajectoires de (4.4) partant de z_0 et arrivant proche de z. C'est donc une question de contrôle de trajectoires guidées par le système différentiel (4.9).

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

La contrôlabilité du système précédent sera discuté en détail dans la dernière partie de ce mémoire, mais elle est déjà importante ici. S'il est à peu près clair que partant d'un point z_0 quelconque de $\mathbb{R}^d \times \mathbb{R}^d$ on puisse atteindre un point quelconque sur la coordonnée x, on ne peut pas en dire autant pour la coordonnée y puisque la présence de φ n'impacte directement que $x(t)$ et non $y(t)$. Notamment, il faut revenir au problème initial et remarquer que $y(t)$ s'écrit

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

De ce fait, si ∇U est borné, y ne peut atteindre des valeurs au delà de $B(y_0, \|\nabla U\|_\infty)$ et l'examen de (4.10) incite donc à supposer que ∇U est une application surjective de $\mathbb{R}^d$ dans $\mathbb{R}^d$. Ainsi, nous démontrons alors la proposition suivante exploitant cette idée.

Proposition 4.2.3 *Si l'hypothèse* $(\mathbf{H_{Hypo}})$ *est vraie et si de plus* $\lim_{|x|\to+\infty} \frac{U(x)}{|x|} = +\infty$, *alors les deux points suivants sont vrais.*
(i) Pour tout $T > 0$, *pour tout* $z_0 \in \mathbb{R}^d \times \mathbb{R}^d$ *et pour tout* $\mathcal{O} \subset \mathbb{R}^d \times \mathbb{R}^d$, $P_T(z_0, \mathcal{O}) > 0$. *Ainsi, pour tout* $z_0 \in \mathbb{R}^{2d}$, $p_T(z_0, .)$ *est* λ_{2d}*- p.s. positif et il existe au plus une mesure invariante pour* $(X_t, Y_t)_{t\geqslant 0}$ *dans le cas où* $r(t) \longmapsto r_\infty \in (0; +\infty)$.
(ii) Si de plus r *est constant positif et qu'il existe un minimum* x^* *de* U *tel que* $D^2U(x^*)$ *est inversible, alors en notant* $z^* = (x^*, 0)$, *il existe* $T > 0$ *tel que pour tout compact* K *de* $\mathbb{R}^{2d}$, *il existe* $\nu_K > 0$ *et* $\alpha(T, K) > 0$ *pour lesquels*

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

Le premier point de la proposition précédente exploite la contrôlabilité du système (4.9) en utilisant la transformée de Fenchel-Legendre de U : si $\lim_{|x|\to+\infty} \frac{U(x)}{|x|} = +\infty$, alors pour un ν quelconque de $\mathbb{R}^d$ l'application $F_\nu(x) = \langle \nu, x\rangle - U(x)$ possède un maximum et par conséquent ∇U est surjectif. Pour atteindre n'importe quel ouvert $\mathcal{O}$ de $\mathbb{R}^d \times \mathbb{R}^d$, l'idée est alors de fabriquer une trajectoire en trois parties : la première partie de la trajectoire amène rapidement x en un point $x(\eta)$, la seconde partie de la trajectoire reste constante en x entre η et $T-\eta$ puis la dernière partie de la trajectoire amène x dans $\Pi_x(\mathcal{O})$. Bien entendu, $x(\eta)$ est calibré pour que le temps passé entre η et $T - \eta$ sur ce point permette à $y(T - \eta)$ d'arriver dans $\Pi_y(\mathcal{O})$ et l'existence d'un

tel $x(\eta)$ est assurée par la surjectivité de ∇U. Cette stratégie démontre ainsi la contrôlabilité approchée du système (4.9).

Le second point de la proposition revêt une importance cruciale pour obtenir que les compacts sont des *petite sets* pour appliquer ensuite les estimées de Meyn et Tweedie. La proposition s'appuie donc sur un résultat plus fort sur le système (4.9) de contrôlabilité locale exacte autour du point z^*. Le rôle de l'inversibilité de $D^2U(x^*)$ est d'apporter une condition de plein rang de Kalman pour le système linéarisé autour de z^* (on pourra consulter [Coron, 2007] pour plus de détails). Elle pourrait être substituée par toute autre condition qui assure la contrôlabilité exacte localement autour de z^* mais cela semble être géométriquement une condition satisfaisante pour garantir un tel résultat. Cette contrôlabilité exacte permet alors d'obtenir de la masse autour du point z^* et d'y minorer localement la densité p_t, pour ce faire nous utilisons un résultat récent de [Delarue and Menozzi, 2010].

4.2.3 Régime d'équilibre ($r_\infty > 0$)

Mémoire courte portée Nous décrivons dans cette partie les résultats concernant le comportement asymptotique de (X_t, Y_t) dans la situation où la mémoire (contenue dans la fonction $t \mapsto r(t)$) n'est pas trop longue. Nous identifions une telle situation par le biais de la limite de r en $+\infty$. Ce régime d'équilibre correspond au cas où $r(t) \longmapsto r_\infty \in]0, +\infty]$. Nous faisons donc l'hypothèse suivante sur r.

Hypothèse 8 ($\mathbf{H_r}$) *L'application r a une limite strictement positive r_∞ en $+\infty$ (éventuellement $r_\infty = +\infty$). Par ailleurs, on suppose que r varie assez lentement en $+\infty$:*

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

On peut citer deux situations typiques :
- $k(t) = \exp(\lambda t)$ et dans ce cas $r(t) = r_\infty = \lambda$ avec $(X_t^z, Y_t^z)_{t\geqslant 0}$ Markov homogène.
- $k(t) = \exp(t^\alpha)$ avec $\alpha > 1$ et dans ce cas $r_\infty = \lim_{t\to+\infty} r(t) = +\infty$.

Fonction de Lyapunov La stationnarité du processus est alors assurée dès lors que ∇U possède suffisament de force pour rappeler le processus et assurer alors une propriété de tension. L'hypothèse un peu technique est la suivante.

Hypothèse 9 ($\tilde{\mathbf{H}}_\mathbf{U}$) *Il existe $m \in (0, r_\infty)$ et $\varepsilon \in (0, r_\infty - m)$ tel que*

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

$(\tilde{\mathbf{H}}_\mathbf{U})$ est une hypothèse a priori plus forte que $(\tilde{\mathbf{H}}_\mathbf{U})$ mais n'est tout de même pas trop restrictive. Si σ est indépendant de x, elle est satisfaite pour des potentiels tels que $U(x) \sim_{|x|\to+\infty} |x|^q$ dès que $q > 0$ ou même pour $U(x) \sim_{|x|\mapsto+\infty} \ln(|x|+1)^\beta$ avec $\beta > 1$. Si σ varie, il est nécessaire que celui ci ne soit pas trop grand pour $x \mapsto \infty$:
- Pour des croissances polynomiales de U : $U(x) \sim_{|x|\to+\infty} |x|^q$ avec $q > 0$, l'hypothèse est vraie dès que $\|\sigma(x)\sigma^*(x)\| = o(|x|^{q\wedge 2})$ pour $|x| \to +\infty$.
- Pour des croissances logarithmiques U : $U(x) \sim_{|x|\to+\infty} \ln(|x|+1)^\beta$ avec $\beta > 1$, l'hypothèse est satisfaite si $\|\sigma(x)\sigma^*(x)\| = o(\ln(|x|+1)^{\beta-1})$ pour $|x| \to +\infty$.

La clef de l'étude repose alors sur la construction d'une fonction de Lyapunov permettant de contrôler le système dynamique à la fois sur la coordonnée x et y. Il faut noter que la fonction $\mathcal{E}$ définie précédemment par $\mathcal{E}(x,y,t) = U(x) + \frac{|y|^2}{2r(t)}$ n'est pas suffisante ici car elle ne permet d'obtenir des informations que sur la coordonnée y puisque

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

En revanche, il est possible d'aménager la fonction précédente en exploitant son contrôle de la coordonnée y en y adjoignant un terme croisé « position - vitesse »pour obtenir également des informations sur la position x. On choisit ainsi un couple $(m_\varepsilon, \varepsilon)$ par le biais de $(\tilde{\mathbf{H}}_\mathbf{U})$ et on pose

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

où l'application ρ est définie par

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

Cette fonction V permet d'obtenir une vraie force de rappel sur les coordonnées x et y puisqu'on peut montrer que pour t assez grand

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

Mesures d'occupation Pour $z_0 \in \mathbb{R}^d \times \mathbb{R}^d$, on considère les deux familles de mesures d'occupation $(\nu_t^{z_0}(\omega, dx, dy))_{t\geqslant 1}$ et $(\mu_t^{z_0}(dx,dy))_{t\geqslant 1}$ définies ainsi : si f désigne une fonction continue bornée $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, on pose :

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

et

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

On peut démontrer le premier résultat d'ergodicité sur $(\mu_t^{z_0})_{t\geqslant 0}$:

Théorème 4.2.1 *Supposons $(\tilde{\mathbf{H}}_\mathbf{U})$ et $(\mathbf{H_r})$ avec $r_\infty \in \mathbb{R}_+^* \cup \{+\infty\}$, pour tout $z_0 =\in \mathbb{R}^d \times \mathbb{R}^d$, $(\mu_t^{z_0})_{t\geqslant 1}$ est tendue. Si μ_∞ est un point d'accumulation de $(\mu_t^z)_{t\geqslant 1}$ lorsque $t \to +\infty$:*

(i) Si $r_\infty = +\infty$, la première marginale (en x) de μ_∞ est une mesure invariante du processus de Kolmogorov

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

(ii) Si $r(t) \xrightarrow{t\to+\infty} r_\infty < +\infty$, μ_∞ est une distribution invariante du processus de Markov homogène solution de (4.4) avec $r(t) = r_\infty$, $\forall t \geqslant 0$.

Il est également possible d'obtenir un résultat de convergence des mesures d'occupation aléatoires $(\nu_t^{z_0}(\omega, dx, dy))_{t\geqslant 1}$ sous des hypothèses un petit peu plus forte.

Hypothèse 10 $(\check{\mathbf{H}}_\mathbf{U})$ *Il existe $a \in (0,1]$, $\beta \in \mathbb{R}$ et $\alpha > 0$ tels que*

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

Cette condition et les résultats qui en découlent sont très comparables à ceux du théorème précédent et sont détaillés dans [16].

Mesures stationnaires et Vitesses de convergence Il est possible de décrire la nature de l'équilibre pour le processus (X_t, Y_t) dans le cas où $0 < r_\infty < +\infty$.

Proposition 4.2.4 *Supposons que* $(\mathbf{H_r})$, $(\mathbf{H_{Hypo}})$ *et* $(\tilde{\mathbf{H}}_\mathbf{U})$ *soient satisfaites et que* $r(t) = r_\infty \in \mathbb{R}_+^*$. *Si* $\lim_{|x|\to+\infty} \frac{U(x)}{|x|} = +\infty$, *alors il existe une unique mesure invariante* ν *vérifiant*

- *(i)* ν *est absolument continue par rapport à la mesure de Lebesgue de densité notée* p_{r_∞} *qui est* $\mathcal{C}^\infty(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}_+)$. *De plus,* p_{r_∞} *est l'unique mesure de probabilités solution positive de*

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

- *(ii) Si* $d = 1$, $U(x) = x^2/2$ *avec* $\sigma(x) = \sigma > 0\ \forall x \in \mathbb{R}$, *et* $r(t) = r_\infty \in]0; +\infty[$, *alors* p_{r_∞} *est une mesure Gaussienne centrée de matrice de covariance*

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

Remarque 4.2.1 *Si la situation est simple lorsque* $r_\infty = +\infty$ *puisque le théorème 4.2.1 rapproche le comportement de notre diffusion moyennée à celle de Kolmogorov, dans le cas* $r_\infty \in (0, +\infty)$ *la mesure stationnaire limite est clairement non standard puisque même dans le cas gaussien, la densité* p_{r_∞} *est celle d'une gaussienne « tordue ». Plus* r_∞ *est petit et plus la mémoire est longue portée, ce qui rend la variance de* p_{r_∞} *explosive. Enfin, dans le cas général, l'équation aux dérivées partielles satisfaite par* p_{r_∞} *ne se résout pas et une forme analytique de* p_{r_∞} *reste donc complexe à établir.*

Enfin, il est possible de donner des résultats de convergence de $P_t(z_0, .)$ lorsque $t \to +\infty$. Dans le cas homogène $r(t) = r_\infty$, il est possible d'utiliser des techniques issues des travaux de [Down et al., 1995].

Théorème 4.2.2 *Supposons que la mémoire* r *soit homogène :* $r(t) = r_\infty > 0$ *et que les conditions d'hypo-ellipticité de la proposition 4.2.3, ii) soient vraies. Si* U *vérifie* $(\check{\mathbf{H}}_\mathbf{U})$ *pour un certain* $a \in (0, 1]$, *alors pour tout* $p \geqslant 1$ *et tout* $t \geqslant 0$:

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

où $z = (x, y)$, V_∞ *est une fonction positive donnée par* $V_\infty(z) = U(x) + \frac{r_\infty}{2}\left|x - \frac{y}{r_\infty}\right|^2$, γ_{p,r_∞} *et* C_{a,p,r_∞} *sont des constantes positives explicites qui ne dépendent pas de* z_0 *et* t.

Notons qu'il est également possible de donner des résultats de vitesse de convergence lorsque $r_\infty = +\infty$ en utilisant des arguments de couplage à la diffusion $dX_t = -\nabla U(X_t)dt + \sigma(X_t)dB_t$. Nous renvoyons à [16] pour plus de détails.

4.2.4 Régime explosif ($r_\infty = 0$)

Nous étudions enfin la stabilité du système (4.4) lorsque la mémoire r est à longue portée, ceci est traduit dans notre cas par $\lim_{t\to+\infty} r(t) = r_\infty = 0$. Le cas typique de telles mémoires est le cas où $k(t) = (1 + t)^\alpha$ pour $\alpha > 0$ ou bien $k(t) = e^{(1+t)^\alpha}$ avec $0 < \alpha < 1$.

Potentiel sous-quadratique Nous avons obtenu un résultat assez précis dans le cas sous-quadratique pour le potentiel U. Ce résultat est synthétisé dans le théorème suivant.

Théorème 4.2.3 *Supposons qu'il existe* C *tel que* $|\nabla U|^2 \leqslant C(1+U)$ *et* $\lambda_0 > 0$ *pour lequel* $\mathrm{Tr}(\sigma^* D^2 U \sigma)(x) \geqslant \lambda_0 > 0$. *Si* $r_\infty = 0$ *et pour* t *assez grand* $r'(t) + 2r^2(t) \geqslant 0$, *alors quel que soit* z_0

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

En particulier, il existe une suite $(t_n)_{n\geqslant 1}$ *telle que* $\mathbb{E}[|X_{t_n}^{z_0}|^2] \to +\infty$.

Ce théorème s'applique en particulier pour le cas où la pondération est uniforme sur tous les temps passés avant t : $Y_t = \frac{1}{1+t}\int_0^t \nabla U(X_s)ds$. Ce résultat nous incite donc à penser que si l'on utilise une dynamique moyennée avec une mémoire longue portée, il convient de diminuer d'autant la volatilité pour rendre le système dynamique stable. Ce phénomène s'explique par analogie avec le système physique (HBF) décrit en introduction puisqu'on peut ramener par un changement de temps adéquat notre diffusion à une perturbation diffusion de (HBF). Pour plus de détails, on pourra se rapporter à l'introduction de [16].

Cas quadratique Il est possible d'être extrêmement précis lorsqu'on suppose que U est quadratique. Au vu du résultat obtenu dans la proposition 4.2.4 ii), on constate que lorsque $r_\infty \mapsto 0$, la matrice de covariance devient « infinie »sur la coordonnée x et comme le processus est $(X_t, Y_t)_{t\geqslant 0}$ est gaussien, sa mesure stationnaire si elle existe ne peut qu'être gaussienne, ce qui amène à une infirmation de son existence.

Plus précisément, nous supposons que $U(x) = x^2/2$, $d = 1$ et que la mémoire est polynomiale : $k(t) = (1+t)^\alpha$ (ainsi, $r(t) = \alpha/(1+t)$). Toute l'information est alors contenue dans la donnée de $f(t) = \mathbb{E}[X_t^2]$, $g(t) = \mathbb{E}[Y_t^2]$ et $h(t) = \mathbb{E}[X_t Y_t]$. La formule d'Itô montre que

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

On montre alors le résultat suivant.

Théorème 4.2.4 *Si* $d = 1$, $U(x) = x^2/2$ *et* $k(t) = (1+t)^\alpha$ *avec* $\alpha > 1/2$, *on a :*

i) Quelque soit z_0, $(X_t^{z_0}, Y_t^{z_0})_{t\geqslant 0}$ *est asymptotiquement centré.*

ii) Le processus $(X_t^{z_0}, Y_t^{z_0})_{t\geqslant 0}$ *satisfait*

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

iii)

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

4.3 Cas particuliers d'équations de Fokker-Planck cinétiques

4.3.1 Modèle

Les propriétés intéressantes pour les diffusions moyennées (4.4) sont naturellement celles qui guident les applications en optimisation donc le comportement lorsque le paramètre de diffusion est petit ainsi que la décroissance de la norme $\|P_t - \mu_\infty\|_{\mathbb{L}^2(\mu_\infty)\circlearrowleft}$. Plus précisément, il est important de bien connaître les constantes intervenant dans les décroissances exponentielles notamment pour effectuer un recuit simulé. Les résulats obtenus dans [16] ne sont pas aussi bon qu'espérés puisque nous avons des résultats uniquement en variation totale et les constantes intervenant dans les vitesses ne sont pas forcément optimales avec cette équation cinétique particulière. Nous étudions dans [15] une situation mieux connue qui est l'équation cinétique de Fokker-Planck. Ces équations sont définies au travers du semi-groupe $(P_t)_{t\geqslant 0}$ donné par

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

où a est un paramètre positif et W_t à nouveau un mouvement brownien standard. S'il n'est pas possible de trouver un changement de variable linéaire qui permette de passer de (4.4) à (4.13) [4], ces équations possèdent du moins visuellement des similitudes fortes. J'ai donc été naturellement amené à considérer le calcul le plus précis possible des normes $\|P_t - \mu_\infty\|_{\mathbb{L}^2(\mu_a)\circlearrowleft}$ où μ_a désigne la mesure stationnaire de (4.13) qui est ici explicite (contrairement à celle de (4.4)).

4.3.2 Calcul de la norme $\mathbb{L}^2(\mu_a)_\circlearrowleft$ pour $U = 0$.

Dans [15], nous donnons des résultats exacts de calcul de norme $\mathbb{L}^2(\mu_a)_\circlearrowleft$ pour le semi-groupe de Fokker-Planck cinétique dans des cas très particuliers.

Le premier modèle est réduit à $\mathbb{T} \times \mathbb{R}$ pour les coordonnées "position $\times$ vitesse" où $\mathbb{T} := \mathbb{R}/(2\pi\mathbb{Z})$. Étant donné $a > 0$, l'opérateur est alors

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

qui correspond au cas particulier initié à partir de (4.13) lorsque $U = 0$ et que l'espace d'état en x est compact. Il est facile de voir que $P_t^{(a)}$ converge vers $\mu_a = \lambda \otimes \gamma_a$ où λ est la mesure uniforme sur $\mathbb{T}$ et $gamma_a$ est la distribution gaussienne centrée de variance a. Il est possible d'obtenir après des calculs relativement techniques le premier calcul.

Théorème 4.3.1 *Pour tout $a > 0$ et $t \geqslant 0$, on a*

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

où $\|.\|_{\mathbb{L}^2(\mu_a)\circlearrowleft}$ est ici la norme opérateur dans $\mathbb{L}^2(\mu_a)$.

La technique de preuve consiste à décomposer assez naturellement le générateur L_a sur une base de $\mathbb{L}^2(\mu_a)$ obtenue par tensorisation de la base des polynômes trigonométriques en x et des polynômes d'Hermite en y. Nous identifions alors des sous-espaces orthogonaux (de dimension infinie) stables par L_a, notés formellement $\mathcal{V}_{p_{p\geqslant 0}}$ et sur lesquels L_a agit sous la forme d'une matrice tridiagonale antisymétrique. À noter que c'est précisément cette nature antisymétrique

4. Dans le cas Gaussien, on peut au moins reparamétrer (4.4) en $dX_t = Y_t dt$ et $dY_t = -(X_t + Y_t)dt + dW_t$

qui rend complexe le calcul de la norme opérateur de L_a puisque les vecteur propres ne sont alors plus orthogonaux.

La clef pour identifier les vecteurs propres et valeurs propres de L_a sur chaque $\mathcal{V}_p$ est de décomposer l'opérateur en $D + c_{a,p}S - c_{a,p}S^\star$ et utiliser alors l'algèbre de Lie engendrée par D, S et $S^\star$ qui est de dimension 3. C'est cette propriété fondamentale qui permet le calcul complet du spectre de L_a dans ce cas précis qui est réel pour toute valeur de a, ainsi que de tous les vecteurs propres associés. Plutôt que de détailler les calculs extrêmement techniques[5], attardons-nous sur les conclusions numériques portées par le théorème 4.3.1.

4.3.3 Comportement qualitatif, $U = 0$

Il peut être intéressant de regarder le comportement de la norme décrite dans le théorème 4.3.1 pour les temps petits et grands. Lorsque $t \mapsto 0_+$,

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

Ceci montre que la norme décroît lentement initialement et la puissance 3 doit être considérée comme le premier ordre d'hypocoercivité de l'opérateur L_a. Par ailleurs, lorsque t tend vers $+\infty$,

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

qui illustre la convergence à l'équilibre du semi-groupe $(P_t^{(a)})_{t\geqslant 0}$. Cette borne est bien entendu cohérente avec les bornes générales obtenues sur le système de Fokker-Planck cinétique mais sont en réalité plus explicites. En effectuant une réduction d'échelle en temps et position, on peut déduire du théorème 4.3.1 un corollaire :

Corollaire 4.3.1 *Pour tous $a, c > 0$ et $b \in \mathbb{R}\backslash\{0\}$, si on définit*

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

qui admet $\mu_{a/c}$ comme mesure invariante, alors le semi-groupe $(P_t^{(a,b,c)})_{t\geqslant 0}$ vérifie

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

En particulier, la vitesse de convergence exponentielle est donnée par

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

Il est naturel de comparer cette vitesse avec celle du semi-groupe du noyau de la chaleur $(Q_t^{(a)})_{t\geqslant 0}$ sur $\mathbb{T}$ généré par l'opérateur $K_a := a\partial_x^2$ qui transmet à chaque temps la même quantité a d'aléas que le générateur $L_{a,b,c}$ (où $b \in \mathbb{R}$ et $c > 0$ sont des paramètres que l'on peut régler librement). Comme K_a est auto-adjoint dans $\mathbb{L}^2(\lambda)$ et admet a comme trou spectral, on a

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

Ainsi,si l'on devait choisir entre une procédure de Monte Carlo basée $(Q_t^{(a)})_{t\geqslant 0}$ ou une autre utilisant $(P_t^{(a,b,c)})_{t\geqslant 0}$ pour simuler la mesure uniforme λ, il serait plus efficace de choisir la

5. À la limite de l'*overkill* selon certains relecteurs !

simulation hypo-coercive avec $c > a$ et $b/c > 1$ et projeter sur la première coordonnée. Bien entendu, cet exemple est uniquement illustratif puisque simuler un mouvement Brownien est plus complexe que simuler une loi uniforme sur $\mathbb{R}$ et montre que des convergences à l'équilibre peuvent être plus rapides pour des processus non réversibles que pour des générateurs peut être plus naturel en apparence et réversibles. Les travaux de [Diaconis et al., 2010a] ont montré l'existence de situations similaires avec des chaines de Markov du second ordre.

4.3.4 Étude du processus Ornstein-Uhlenbeck hypocoercif.

Nous étudions ensuite comme second modèle le cas d'un processus d'Ornstein-Uhlenbeck hypocoercif défini sur $\mathbb{R} \times \mathbb{R}$ par le biais de

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

Là encore, la mesure stationnaire est explicite donnée par $\tilde{\mu}_a := \gamma_{1/a} \otimes \gamma_1$. Nous allons donc étudier l'évolution du semi-groupe $(\tilde{P}_t^{(a)})_{t\geqslant 0}$ dans $\mathbb{L}^2(\tilde{\mu}_a)$. Là encore, l'idée est d'étudier l'effet de $\tilde{L}_a$ sur la base de $\mathbb{L}^2(\tilde{\mu}_a)$ obtenue par tensorisation des polynômes d'Hermite en x et y. Nous identifions à nouveau des sous-espaces orthogonaux stables par $\tilde{L}_a$ notés formellement $\tilde{\mathcal{V}}_p$ ici et sur lesquels l'opérateur possède une décomposition similaire à ce qui est déjà rencontré dans le cas $U = 0$, à savoir $\tilde{D} + \tilde{c}_{a,p}\tilde{S} - \tilde{c}_{a,p}\tilde{S}^\star$, ce qui permet à nouveau l'étude complète du spectre de $\tilde{L}_a$ sur chaque $\tilde{\mathcal{V}}_p$ puis dans $\mathbb{L}^2(\tilde{\mu}_a)$. Cette fois, le spectre change de nature selon que a est inférieur ou non à $1/4$: si $a < 1/4$, le spectre est réel et $\tilde{L}_a$ est diagonalisable dans une base (non orthonormée) de $\mathbb{L}^2(\tilde{\mu}_a)$. Si $a > 1/4$, la même propriété est vraie mais le spectre n'est plus réel. Enfin, si $a = 1/4$, l'opérateur n'est plus diagonalisable et possède des blocs de Jordan de tout ordre. On peut enfin étudier l'évolution de la norme de $\tilde{P}_t^{(a)} - \tilde{\mu}_a$ dans $\mathbb{L}^2(\tilde{\mu}_a)$.

Théorème 4.3.2 *Pour tout $a > 0$ et $t \geqslant 0$, on a*

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

où $\|.\|$ désigne la norme opérateur dans $\mathbb{L}^2(\tilde{\mu}_a)$ et $C_a(t)$ est donné par

- *Si $a \in (0, 1/4)$, on note $\theta := \sqrt{1-4a}$ et*

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

- *Si $a \in (1/4, +\infty)$, on note $\theta := \sqrt{4a-1}i$ et*

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

- *Si $a = 1/4$,*

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

Là encore, si $t > 0$ tend vers 0, nous obtenons une décroissance de l'ordre de t^3 (voir [15] pour les calculs précis) alors que lorsque $t \mapsto +\infty$, la décroissance exponentielle est différente selon que a est plus grand ou plus petit que $1/4$. Il convient de retenir que si $a > 1/4$, la fonction

$C_a(t)$ est oscillante et possède une période $T_a = 2\pi/\sqrt{4a-1}$, ce qui entraine alors une pente nulle dans la décroissance de $\|\tilde{P}_t^{(a)} - \tilde{\mu}_a\|_{\mathcal{L}^2(\tilde{\mu_a}}$ tous les instants $kT_a, k \in \mathbb{N}$. On peut également étendre l'étude précédente au cas du générateur

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

qui admet $\tilde{\mu}_{a,b,c,d} := \gamma_{bc/(ad)} \otimes \gamma_{c/d}$ comme mesure invariante. On obtient comme hypocoercivité

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

On peut comparer cette décroissance avec celle du semi-groupe $(\tilde{Q}_t^{(a,b,c,d)})_{t\geqslant 0}$ engendré par $\tilde{K}_{a,b,c,d} := c\partial_x^2 - \frac{da}{b}x\partial_x$ qui injecte la même quantité d'aléas par unité de temps que $\tilde{L}_{a,b,c,d}$ et qui est réversible pour sa mesure invariante $\gamma_{bc/(ad)}$ (qui est bien sûr la marginale en x de $\tilde{\mu}_{a,b,c,d}$). Après changement d'échelle en temps et en espace, on constate que $\tilde{K}_{a,b,c,d}$ est un processus d'Ornstein-Ulhenbeck dont le générateur a pour trou spectral da/b. Ainsi, la décroissance exponentielle de $(\tilde{Q}_t^{(a,b,c,d)})_{t\geqslant 0}$ vers sa mesure stationnaire $\gamma_{bc/(ad)}$ est da/b. Si on choisit

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

on constate qu'il est plus efficace d'utiliser la première composante du semi-groupe $(\tilde{P}_t^{(a,b,c,d)})_{t\geqslant 0}$ plutôt que $(\tilde{Q}_t^{(a,b,c,d)})_{t\geqslant 0}$ pour simuler $\gamma_{bc/(ad)}$. On obtient ainsi la même conclusion que dans le paragraphe précédent.

4.4 Diffusion moyennée à petit paramètre

Nous revenons désormais au cas de la diffusion à gradient moyenné (4.4) par l'étude des petites perturbations aléatoires de ce système dynamique. Nous abordons notamment cette problématique en étudiant la situation trajectorielle ainsi que la question plus délicate des mesures invariantes. Dans cette étude, nous nous restreignons volontairement à la situation Markov homogène correspondant au cas où la fonction de mémoire est $k(t) = e^{\lambda t}$ dans les paragraphes précédents :

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

Rappelons que z désignera un couple (x, y) tandis que $(Z_t^\varepsilon)_{t\geqslant 0}$ sera le processus couplé $(X_t^\varepsilon, Y_t^\varepsilon)_{t\geqslant 0}$ avec un niveau ε de bruit associé à (4.20). Dans la suite, on notera ν_ε l'unique mesure stationnaire de (4.20) (sous l'hypothèse 7 notée $\mathbf{H_{Hypo}}$) ainsi que $(P_t^\varepsilon(z,.)$ le semi-groupe associé à (4.20) et enfin $\mathcal{A}^\varepsilon$ son générateur.

4.4.1 Grandes déviations trajectorielles

Bien entendu, l'étude limite de (4.20) lorsque $\varepsilon \to 0$ est intimement au comportement du système dynamique déterministe obtenu pour $\varepsilon = 0$ qui est défini par

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

Le lien sera déduit entre autres choses des solutions "optimales" au problème de contrôle déjà mentionné pour obtenir la minoration de la densité $p_t(z_0, z)$. Si on définit sur $\mathbb{R}^d \times \mathbb{R}^d$ le champ de vecteur de drift $b(z) = (-y, \lambda[\nabla U(x) - y])$, le système contrôlé consiste à étudier étant donnée $\varphi \in \mathbb{H}_0^1$ (espace de Cameron-Martin) le comportement de $\mathbf{z}_\varphi := (\mathbf{z}_\varphi(t))_{t\geqslant 0}$ et $\tilde{\mathbf{z}}_\varphi := (\tilde{\mathbf{z}}_\varphi(t))_{t\geqslant 0}$, vérifiant

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

Notons que sous l'hypothèse 10 notée $(\tilde{\mathbf{H}}_\mathbf{U})$ dans la section 4.2, on peut démontrer que les trajectoires contrôlées n'explosent pas en temps fini. Par ailleurs, nous établissons dans un premier temps un Principe de Grandes Déviations (P.G.D.) en temps fini, malgré la dégénerescence du mouvement brownien sur y.

Proposition 4.4.1 *Supposons que U satisfasse la condition $(\tilde{\mathbf{H}}_\mathbf{U})$, alors pour tout $z \in \mathbb{R}^d$ et toute suite $(z_\varepsilon)_{\varepsilon>0} \to z$ lorsque $\varepsilon \to 0$, le processus $Z^\varepsilon = (X^{(\varepsilon)}, Y^{(\varepsilon)})$ satisfait un P.G.D.sur $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^{2d})$ (muni de la topologie associée à la convergence uniforme sur tout compact). Son taux est ε^{-2} et la (bonne) fonction de taux $\mathcal{I}_z$ est définie pour toute fonction absolument continue $\mathbf{z} = (\mathbf{z}(t))_{t\geqslant 0}$ vérifiant $\mathbf{z}(0) = z$ par*

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

En particulier, pour tout $t \geqslant 0$ et $z \in \mathbb{R}^{2d}$, $(P_t^\varepsilon(z_\varepsilon, .))_{\varepsilon>0}$ satisfait un P.G.D. de vitesse ε^{-2} et de fonction de taux $I_{z,t}$ définie pour tous $z, z' \in \mathbb{R}^{2d}$ par

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

où $\mathcal{Z}_t(z, z')$ désigne les trajectoires absolument continues $\mathbf{z}$ menant z à z' en un temps t.

La difficulté principale de cette proposition consiste à établir le principe de contraction qui est intimement lié à la propriété de non-explosion des trajectoires contrôlées. Cette non-explosion en temps fini se démontre en établissant une perturbation du lemme de Gronwall sur la fonction (de Lyapunov) $\mathcal{E}(x, y) = U(x) + |y|^2/(2\lambda$. Remarquons que c'est également un cas d'application des extensions du théorème de Schilder démontrées dans [Azencott, 1980].

4.4.2 Grandes déviations extraites de $(\nu_\varepsilon)_{\varepsilon \mapsto 0}$

La propriété de grande déviation sur les mesures stationnaires $(\nu_\varepsilon)_{\varepsilon \mapsto 0}$ dépend ensuite de plusieurs conditions. La première condition à satisfaire est celle de propriété de tension exponentielle pour la suite de mesures. Cette propriété s'établit en étudiant les moments des temps d'entrée dans les compacts du processus $(Z_t^\varepsilon)_{t\geqslant 0}$ lorsque $\varepsilon \mapsto 0$. Ces temps d'entrée s'étudient en utilisant encore une fois une fonction de Lyapunov. Là encore, pour obtenir cette propriété de tension, la fonction usuelle évoquée précédemment $\mathcal{E}(x, y)$ n'est pas suffisante et il faut utiliser une fonction contrôlant globalement sur les coordonnées x et y de (4.20). L'astuce pour cela consiste à utiliser non pas $V(x, y)$ introduite dans (4.11) mais la fonction

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

et d'opérer un calibrage adéquat des coefficients δ et p afin d'obtenir une propriété de contraction du type

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

Il est alors possible d'obtenir le résultat suivant.

Proposition 4.4.2 *Si* U *satisfait* $(\tilde{\mathbf{H}}_U)$, *alors il existe un compact* B *de* $\mathbb{R}^{2d}$, *tel que le temps d'atteinte* τ_ε *de* B *vérifie :*

i) *Pour tout compact* K, $\limsup_{\varepsilon\to 0}\sup_{z\in K}\mathbb{E}_z[(\tau_\varepsilon)^2] < \infty$.
ii) *Il existe* $\delta > 0$ *tel que pour tout 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) *Pour tout compact* K *tel que* $K \cap B = \varnothing$, $\liminf_{\varepsilon\to 0}\inf_{z\in K}\mathbb{E}_z[\tau_\varepsilon] > 0$.

On peut alors en déduire le résultat principal de cette partie qui établit un P.G.D. extrait et donne une équation de Hamilton-Jacobi vérifiée par toutes les fonctions de taux.

Théorème 4.4.1 *Si* U *satisfait* $(\tilde{\mathbf{H}}_U)$, *alors* $(\nu_\varepsilon)_{\varepsilon\in(0,1]}$ *est exponentiellement tendue. De plus, pour toute sous-suite* $(\varepsilon_n)_{n\in\mathbb{N}}$ *pour laquelle* $(\nu_{\varepsilon_n})_{n\in\mathbb{N}}$ *satisfait un P.G.D.*[6] *de taux* ε_n^{-2}, *la fonction de taux* W *vérifie*

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

Ce théorème ne donne qu'un résultat partiel pour l'existence d'un P.G.D. pour la suite $(\nu_\varepsilon)_{\varepsilon\geqslant 0}$ puisqu'il n'est pas clair que les fonctions de taux obtenues à partir de (4.24) soient toutes identiques. En effet, l'équation d'Hamilton Jacobi (ici présentée sous la forme contrôle optimal de la programmation dynamique) ne possède pas nécessairement une propriété d'unicité de ses solutions assurant alors l'unicité du W limite. L'objectif du paragraphe suivant est de donner des hypothèses suffisantes pour obtenir un tel P.G.D. le long de n'importe quelle sous-suite $(\varepsilon_n)_{n\in\mathbb{N}}$.

4.4.3 Estimées de Freidlin & Wentzell

Équilibres du champ de vecteur L'hypothèse fondamentale à ajouter pour comprendre le système dynamique défini par (4.4) est la suivante.

Hypothèse 11 ($\mathbf{H_D}$) *L'ensemble des points critiques de* U *est discret (et donc fini), et la Hessienne de* U *est inversible en ces points critiques.*

On notera par la suite $\{x_1^\star, \dots x_\ell^\star\}$ l'ensemble des points critiques de U. La propriété élémentaire suivante permet d'identifier les équilibres du système dynamique (4.21) utilisant le champ de vecteur $+b$.

Proposition 4.4.3 *Sous l'hypothèse* $(\mathbf{H_D})$, *les équilibres de* (4.21) *sont les éléments* $\{z_1^\star, \dots, z_\ell^\star\} := \{(x_1^\star, 0), \dots, (x_{ell}^\star, 0)\}$. *Les points stables sont ceux tels que* $x_i^\star$ *est un minimum (local) de* U.

Sous cette hypothèse $(\mathbf{H_D})$, il est possible d'étendre l'équation (4.24) à l'horizon infini, c'est-à-dire toute bonne fonction de taux W est en fait solution de :

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

6. Une telle suite sera dite LD-convergente

La démonstration de ce résultat s'appuie uniquement sur des arguments dynamiques du champ de vecteur $+b$, non explosion des trajectoires en temps infini, compacité des trajectoires et ensembles ω-limite. En particulier, la démonstration de (4.25) ne propose aucun argument pour assurer l'unicité de W. Cependant, la chose importante qu'on peut lire dans cette formule (4.25) est la dépendance exclusive de la fonction W définie sur $\mathbb{R}^d \times \mathbb{R}^d$ à seulement ses valeurs $W(z_i^\star)$ en les points d'équilibres de $+b$. C'est précisément ces valeurs $W(z_i^\star)$ qui sont données par le biais des estimées de Freidlin & Wentzell.

Théorie de Freidlin & Wentzell L'idée est de s'appuyer sur une expression explicite des mesures invariantes $(\nu_\varepsilon)_{\varepsilon \geqslant 0}$ au travers de la chaîne squelette qu'on peut former en utilisant les temps d'atteinte et de sortie dans des voisinages de ces points d'équilibres. Nous étendons cette formule donnée dans [Has'minskii, 1980] (et dont l'utilisation est une des clefs de la théorie de [Freidlin and Wentzell, 1984]) à notre situation hypo-elliptique en utilisant des arguments de contrôle de trajectoire qui assurent que la représentation squelette de ν_ε est *finie*.

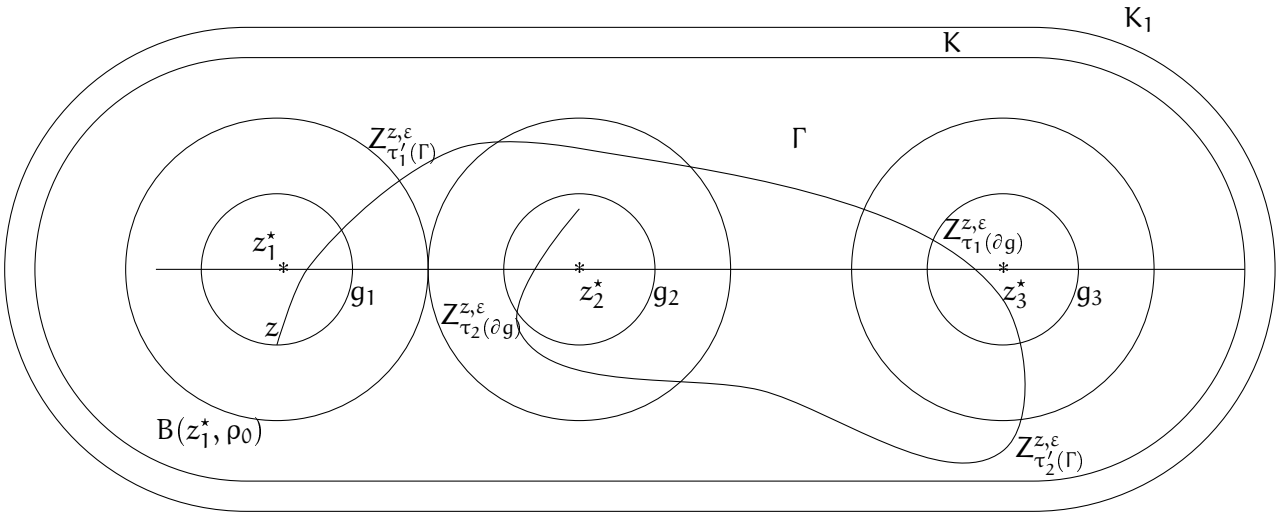


FIGURE 4.1 – Représentation des voisinages g_i des points d'équilibres et du processus $(Z_t^{z,\varepsilon})_{t \geqslant 0}$, la chaîne squelette correspond aux transitions $Z^{z,\varepsilon}_{\tau_1(\partial g)} \mapsto Z^{z,\varepsilon}_{\tau_2(\partial g)}$ qui appartient à $\cup_{i=1}^{\ell} g_i$.

Proposition 4.4.4 *Si $\tilde{\mu}_\varepsilon$ désigne la mesure invariante de la chaîne squelette définie sur $\cup_{i=1}^{\ell} g_i$, alors la mesure définie pour tout borélien A de $\mathbb{R}^d \times \mathbb{R}^d$ par*

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

est une mesure invariante finie proportionnelle à la distribution invariante ν_ε.

Puis nous montrons que les estimées de Freidlin & Wentzell s'appliquent à notre cadre pour la chaîne squelette. On définit pour deux points ξ_1 et ξ_2 le coût de contrôle optimal en temps T

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

et le coût de contrôle optimal $I(\xi_1,\xi_2) := \inf_{T\geqslant 0} I_T(\xi_1,\xi_2)$. De même, on définit

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

Il est possible en utilisant le P.G.D. trajectoriel en temps fini et la contrôllabilité exacte locale au voisinage des $z_i^\star$ d'encadrer les probabilités de transition de la chaîne squelette par I lorsque ε est petit (voir [21] et [Freidlin and Wentzell, 1984]) . Nous démontrons ainsi le résultat suivant

Proposition 4.4.5 *Si on suppose que* U *satisfait les hypothèses* $(\mathbf{H_D})$, $(\tilde{\mathbf{H}}_\mathbf{U})$ *et* $(\mathbf{H_{Hyp}})$, *alors :*

i) *Pour tout couple* $(i,j) \in \{1\dots\ell\}^2$, $\tilde{I}(z_i^\star, z_j^\star) < +\infty$ *et tous les voisinages* g_i *communiquent en une étapge par la chaîne squelette.*

ii) *De plus, pour tout* $\gamma > 0$, *il existe* ρ_0 *et* ρ_1 *(tailles des voisinages des* $z_i^\star$ *dans la chaîne squelette) tels que* $0 < \rho_1 < \rho_0$ *et pour lesquels lorsque* ε *est assez petit on a*

$$\forall (i,j) \in \{1\dots\ell\}^2 \quad \forall x \in \partial g_i \qquad 0 < e^{-\varepsilon^{-2}[\tilde{I}(z_i^\star, z_j^\star)+\gamma]} \leqslant \tilde{\mathbb{P}}^\varepsilon(x, \partial g_j) \leqslant e^{-\varepsilon^{-2}[\tilde{I}(z_i^\star, z_j^\star)-\gamma]}.$$

4.4.4 Principe de Grandes Déviations pour $(\nu_\varepsilon)_{\varepsilon\geqslant 0}$

L'estimation précédente permet de donner un encadrement précis de la mesure stationnaire pour la chaîne squelette au travers de la notion d'{i}-Graphes. Rappelons rapidement que pour tout $i \in \{1,\dots,\ell\}$, on note $\mathcal{G}(i)$ l'ensemble des graphes orientés de sommets $\{z_1^\star,\dots,z_\ell^\star\}$ tels que

(i) Tout sommet $z_j^\star \neq z_i^\star$ est le point de départ d'exactement une arête.

(ii) Le graphe ne contient pas de cycle.

(iii) Pour tout $z_j^\star \neq z_i^\star$, il existe un (unique) chemin composé d'arêtes orientées démarrant à $z_j^\star$ et menant à $z_i^\star$.

Enfin, l'utilisation de l'estimation de μ_ε pour la chaîne squelette déduite de la proposition 4.4.5 conjuguée à la formule de reconstruction de ν_ε donnée dans la proposition 4.4.4 permet alors de conclure le résultat suivant.

Théorème 4.4.2 *Sous les hypothèses* $(\mathbf{H_{Hypo}}), (\mathbf{H_D})$ *et* $(\tilde{\mathbf{H}}_\mathbf{U})$, *pour toute suite* (ε_n) LD-*convergente, la fonction de taux* W *associée vérifie*

$$\forall i \in \{1\dots\ell\} \qquad W(z_i^\star) = \min_{g\in\mathcal{G}(i)} \sum_{(z_m^\star \to z_n^\star)\in g} I(z_m^\star, z_n^\star) = \min_{g\in\mathcal{G}(i)} \sum_{(z_m^\star \to z_n^\star)\in g} \tilde{I}(z_m^\star, z_n^\star). \tag{4.26}$$

Par ailleurs, W *est définie de manière unique explicitement par* (4.26) *et*

$$\forall z \in \mathbb{R}^d \times \mathbb{R}^d \qquad W(z) = \min_{1\leqslant i\leqslant \ell} \inf_{\substack{\varphi \in \mathbb{H}_0^1 \\ \tilde{\mathbf{z}}_\varphi(0) = z,\ \tilde{\mathbf{z}}_\varphi(+\infty) = z_i^\star}} \left[\frac{1}{2}\int_0^\infty |\dot{\varphi}|^2 + W(z_i^\star) \right],$$

et $(\nu_\varepsilon)_{\varepsilon\geqslant 0}$ *satisfait un Principe de Grandes Déviations.*

4.4.5 Quasi-potentiel pour une fonction double-puits

Si l'étude du quasi-potentiel W donné dans le paragraphe précédent dans le théorème 4.4.2 est simple dans le cas où U est convexe, c'est loin d'être dans la situation générale. Nous étudions un cas particulier illustratif pour une fonction définie sur $\mathbb{R}$ présentant un double puits. Typiquement, le potentiel U est ressemble à la fonction donnée dans la figure 4.2.

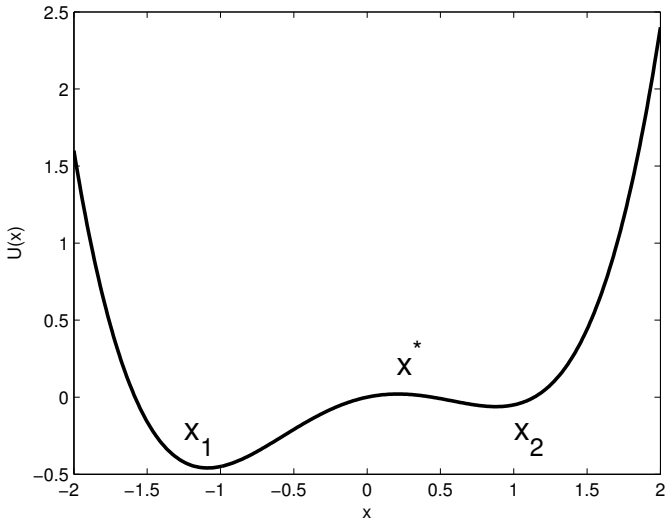


FIGURE 4.2 – Exemple de double puits U avec 2 minima $x_1 < x_2$ et un maximum local x^*.

Il s'agit donc d'évaluer le quasi-potentiel associé au P.G.D. dérivé du processus (4.4), et ici son expression formelle est simple à calculer au sens où seulement les coûts de contrôle L^2 de $z_1^\star := (x_1, 0)$ à $z_2^\star := (x_2, 0)$ sont à évaluer de par la simplicité des i-graphes ici. Sans perte de généralités, on peut supposer $U(x_1) < U(x_2)$ et l'idée est de calculer une minoration de $W(z_2^\star) = I(z_1^\star, z_2^\star)$ ainsi qu'une majoration de $W(z_1^\star) = I(z_2^\star, z_1^\star)$.

Majoration et minoration dans la situation "standard" Dans la situation où l'on considère le problème de contrôle non dégénéré $\dot{z} = -\nabla U(z) + \varphi$, on constate que le choix $\varphi(z) = 2\nabla U(z)$ est un choix qui permet d'amener la trajectoire d'un minimum local x_1 à un maximum local $x^\star$ avec un coût égal à $2[U(x^\star) - U(x_1)]$. Ensuite, un argument de continuité de la fonction de coût permet d'obtenir une trajectoire de coût identique entre x_1 et x_2, ce qui permet d'obtenir facilement que

$$I(x_1, x_2) \leqslant 2[U(x^\star) - U(x_1)].$$

Par ailleurs, un raisonnement simple permet de donner une minoration en remarquant que pour toute trajectoire $(z_t)_{t\geqslant 0}$ menant de x_1 à x_2 passe nécessairement par x^* [7]

$$|\dot{\varphi}|^2 = |\dot{z} + \nabla U|^2 \geqslant 2\langle \dot{z}, \nabla U(z)\rangle.$$

et on minore à nouveau facilement le coût de contrôle par $2[U(x^\star) - U(x_1)]$.

Majoration et minoration dans la situation du gradient moyenné En réalité, on peut étendre ces résultats aux vecteurs de drift un peu plus généraux (voir les travaux de [Sheu, 1986] par exemple) mais le problème est ouvert en ce qui concerne un drift général donné par $b(x, y) = (-y, \lambda(\nabla U(x) - y)$. Afin de trouver une trajectoire permettant de passer de $z_2^\star$ à $z_1^\star$, nous nous inspirons de la méthode utilisée dans le cas standard en choisissant de "remonter" l'équation différentielle dans le temps, ceci se traduisant par une équation différentielle

$$dX_t = \frac{1}{e^{\lambda t}} \int_0^t \lambda e^{\lambda s} \nabla U(X_s) ds. \tag{4.27}$$

7. En dimension supérieure, il faut considérer l'élévation minimale pour traverser la "coline" entre x_1 et x_2

Comme le contrôle φ n'agit que sur la première coordonnée, il est naturel de choisir $\dot{\varphi} = 2y$ qui abouti alors à l'équation différentielle (4.27). Cette méthode permet alors d'exhiber une trajectoire ayant un coût identique au coût optimal dans le cas standard.

Proposition 4.4.6 *Pour un potentiel double-puits décrit précédemment, on a*

$$W(z_1^\star) = I(z_2^\star, z_1^\star) \leqslant 2[U(x^\star) - U(x_2)].$$

La minoration de $W(z_2)$ est nettement plus délicate et peut être abordée en considérant des trajectoires contrôlées dont les contraintes agissent sur x et y, ou bien en se restreignant aux contrôles n'agissant que sur x en gardant en mémoire l'inégalité :

$$I_T(z_1^\star, z_2^\star) = \inf_{\begin{cases}\varphi \in \mathbb{H}_0^1 \\ \mathbf{z}_\varphi(0) = z_1^\star \\ \mathbf{z}_\varphi(T) = z_2^\star\end{cases}} \frac{1}{2}\int_0^T |\dot{\varphi}(s)|^2 ds \geqslant \inf_{\begin{cases}\varphi, \psi \in \mathbb{H}_0^1 \\ \mathbf{z}_{\varphi,\psi}(0) = z_1^\star \\ \mathbf{z}_{\varphi,\psi}(T) = z_2^\star\end{cases}} \frac{1}{2}\int_0^T |\dot{\varphi}(s)|^2 + |\dot{\psi}(s)|^2 ds$$

où $\mathbf{z}_{\varphi,\psi}$ désigne la trajectoire controlée sur x et y par φ et ψ. Nous ne décrirons ici que l'approche utilisant les contrôles "dégénérés" n'agissant que sur x, l'autre méthode se trouvant dans [21] et imposant moins de contraintes sur U mais donnant naturellement de plus mauvaises minorations. Étant donnée une trajectoire controllée par φ quelconque, on écrit que

$$|\dot{\varphi}|^2 = |\dot{x} + y|^2 = \dot{x}^2 + y^2 + 2\dot{x}y,$$

et on cherche à minorer cette forme quadratique en $x, y et \dot{x}$ par la dérivée le long de la trajectoire z_φ d'une fonction de x et y. Ce principe est donc le même que dans l'approche standard lorsqu'on utilisait $|\dot{\varphi}|^2 \geqslant 2\langle \dot{z}, \nabla U(z)\rangle$. Il s'agit donc de trouver une fonction $\mathcal{L}(x, y)$ vérifiant

$$\dot{x}^2 + y^2 + 2\dot{x}y \geqslant \langle \nabla \mathcal{L}(x, y), (\dot{x}, \dot{y})\rangle. \tag{4.28}$$

La fonction $\mathcal{L}$ retenue est alors de la forme

$$\mathcal{L}_{\alpha,\beta,\gamma}(x, y) := \alpha U(x) + \beta y^2/2 - \gamma y U'(x),$$

et il est frappant de constater que cette fonction peut servir à la fois à la compacité en temps long du processus stochastique mais aussi à minorer le coût $\mathbb{L}^2$ de contrôle entre deux points $z_1^\star$ et $z_2^\star$. On obtient alors le résultat suivant.

Proposition 4.4.7 *Pour tout $\alpha \in [0, 2]$, il existe $m(\alpha, \lambda)$ explicite tel que $\|U''\|_\infty \leqslant m(\alpha)$ entraine qu'on peut choisir $\beta(\alpha)$ et $\gamma(\alpha)$ pour que (4.28) soit vraie. Pour un tel choix, on a alors*

$$I_T(z_1^\star, z_2^\star) \geqslant \alpha[U(x^\star) - U(x_1)].$$

On constatera que cette proposition ne permet pas de minorer le coût $I_T(z_1^\star, z_2^\star)$ par un facteur supérieur à 2 devant l'élévation de U, ce qui est cohérent avec le résultat de la proposition 4.4.6. Ces deux dernières propositions associées au théorème 4.4.2 permettent alors d'énoncer le résultat final de concentration de la mesure ν_ε sur le minimum global de U.

Théorème 4.4.3 *Sous les hypothèses $(\mathbf{H_{Hypo}}), (\mathbf{H_D})$ et $(\tilde{\mathbf{H}}_\mathbf{U})$, si U est un potentiel double puits réel décrit plus haut tel que $U(x_1) < U(x_2)$ vérifiant $\|U''\|_\infty \leqslant m\left(2\frac{U(x^\star)-U(x_2)}{U(x^\star)-U(x_1)}, \lambda\right)$, alors*

$$\lim_{\varepsilon \mapsto 0} \nu_\varepsilon = \delta_{x_1}.$$

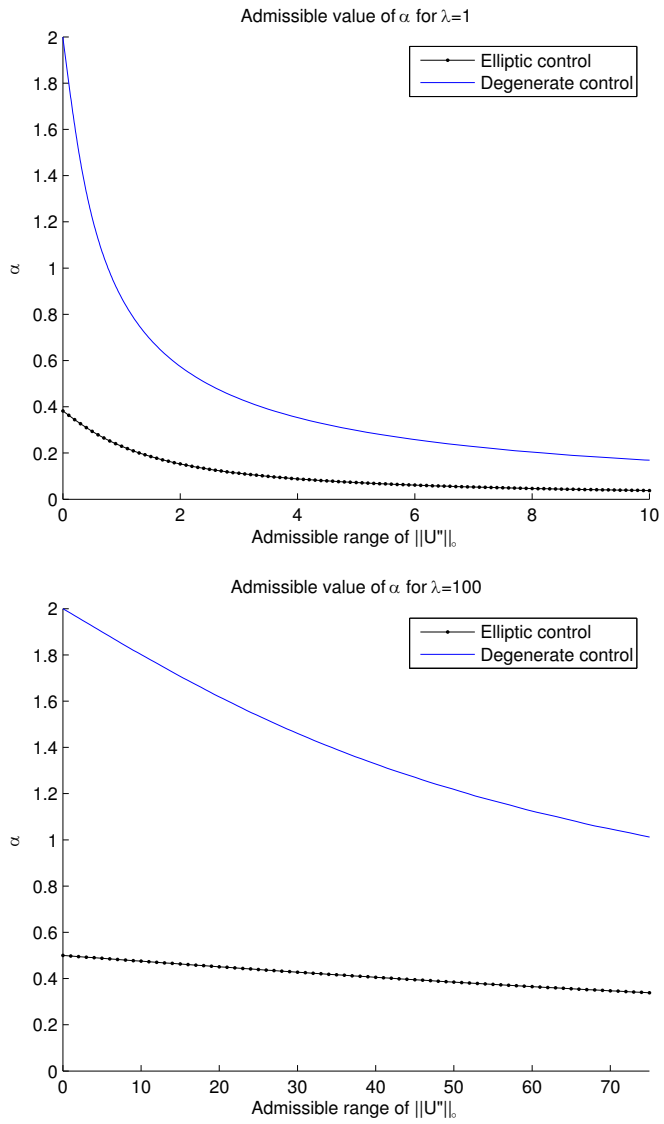


FIGURE 4.3 – Évolution de α coefficient devant l'élévation de U en fonction de $\|U''\|_\infty$ pour différentes valeurs de λ.

Notons que lorsque λ devient grand, le système de gradient moyenné utilise alors une mémoire à plus court terme et la borne sur $\|U''\|_\infty$ dans le résultat précédent devient de plus en plus permissive. On constatera que ce phénomène est illustré par la Figure 4.3 qui présente pour plusieurs valeurs de λ l'évolution de α coefficient devant l'élévation $U(x^\star) - U(x_1)$ en fonction du $\|U''\|_\infty$.

4.5 Élargissements

4.5.1 Hypo-coercivité du gradient moyenné et Recuit simulé

Le premier développement concernant le processus de gradient moyenné auquel on peut penser est la question d'une majoration de la norme dans $\mathbb{L}^2(\nu)$. Ce point est resté en suspens dans l'avancée de mes travaux puisque les résultats énoncés dans ce mémoire ne concerne que des vitesses en variation totale utilisant les techniques de [Down et al., 1995]. Il est tout à fait

envisageable ce résultat à un résultat plus fort en norme $\mathbb{L}^2$ par exemple. Une stratégie pour obtenir ce résultat serait, tout en exploitant un peu plus la fonction de Lyapunov, d'obtenir une inégalité fonctionnelle donnant un lemme de Gronwall, stratégie déjà exploitée dans le cas de l'équation de Fokker-Planck cinétique (voir par exemple [Villani, 2006]). Cette méthode repose sur une inégalité de Poincaré pour la mesure stationnaire qui est explicite dans la situation particulière des équations de Fokker-Planck cinétiques (ce qui n'est pas le cas pour le processus de gradient moyenné sauf dans le cas gaussien).

Après avoir établi un tel résultat d'hypo-coercivité en norme $\mathbb{L}^2$, il serait raisonnable de poursuivre l'étude de ce gradient moyenné en définissant un vrai algorithme de recuit simulé, ε devenant une vraie fonction de t évanescente. Des expérimentations (non présentées ici) ont établi qu'il était possible d'utiliser un schéma de température $\varepsilon(t) = c/\log t$ et d'obtenir une convergence du processus $(Z_t^{\varepsilon(t)})_{t\geqslant 0}$ vers le minimum global du potentiel U dans le cadre d'un potentiel double-puits sous réserve d'une "bosse" suffisament proéminente entre les deux puits (voir condition du théorème 4.4.3). De plus, nous avons plusieurs indices numériques qui laissent penser que la constante c peut être choisie inférieure à la constante limite dans le cadre du recuit simulé standard. Enfin, l'optimisation en λ pour l'algorithme de recuit sur le gradient moyenné semble importante. Tous ces points sont pour le moment sans réponse.

4.5.2 Contrôlabilité du système moyenné

Une autre question intéressante concerne l'étude de la nature exacte des résultats de contrôlabilité qui peuvent être obtenus sur le système 4.22. Nous démontrons dans notre étude que sous des hypothèses de non dégénérescence autour des points critiques de U, et des conditions de croissance en l'infini, la contrôlabilité approchée est vraie. Si la condition de croissance sur U semble inévitable, il n'en est pas de même pour l'hypothèse de non dégénérescence et on pourra consulter [Coron, 2007] pour de nombreux manières de contourner la méthode de Linéarisation de Kalman, telles que les conditions de Sussman, et des méthodes de points fixes peuvent également donner des résultats de contrôlabilité exacte (voir par exemple les travaux de [Beauchard and Zuazua, 2009]).

Enfin, l'implémentation de méthodes numériques permettant de calculer la fonction quasi-potentiel W est encore à faire. Une collaboration naissante avec des spécialistes de théorie du contrôle nous a amené à considérer le principe du maximum de Pontryagin comme brique de base afin d'obtenir des méthodes de simulations.

4.5.3 Simulation par méthodes non réversibles

Les conclusions apportées par le paragraphe 4.3 sont qu'il peut être avantageux d'utiliser des méthodes d'ordre 2 (chaînes de Markov d'ordre 2, équations cinétiques) afin d'obtenir des résultats de convergence à l'équilibre plus rapide qu'une simple dynamique d'ordre 1. Ceci, bien sûr, n'est pas vrai en toute généralité et mériterait d'être étayé au moins par de nombreux exemples génériques, ce qui n'est pas le cas pour le moment en regard des conclusions très partielles obtenues sur le modèle de Fokker-Planck cinétique.

Si un tel phénomène était avéré, il aurait alors de nombreuses conséquences intéressantes pour des algorithmes stochastiques utilisant une phase de simulations de lois, notamment les algorithmes bayésiens et autres méthodes d'optimisation stochastique utilisant une phase de simulation Monte-Carlo.

Bibliographie personnelle

Thèse

[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.

Articles publiés ou acceptés - Statistiques en grandes 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.

Articles publiés ou acceptés - Modèles Déformables

[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.

Articles publiés ou acceptés - Systèmes Dynamiques

[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 L. Miclo. Spectral decompositions and l^2-operator norms of toy hypocoercive models. *Kinetic and Related Models, to appear*, 2012.

[16] S. Gadat and F. Panloup. Long time behavior and stationary regime of memory gradient diffusions. *en révision pour Annales de l'Institut Henri Poincaré (B)*, pages 1–40, 2012.

Prépublications

[17] J. Bigot, S. Gadat, T. Klein, and C. Marteau. Intensity estimation of non-homogeneous poisson processes from shifted trajectories. *Preprint*, 2011.

[18] D. Bontemps and S. Gadat. Bayesian posterior consistency in the functional randomly shifted curves model. *Preprint*, 2012.

[19] C. Cierco, M. Champion, S. Gadat, and M. Vignes. A boost-boost algorithm for high dimensional multivariate regression. *Preprint*, 2012.

[20] C. Cierco, M. Champion, S. Gadat, and M. Vignes. Gene network recovery and $\mathbb{L}^2$ boosting algorithm. *Preprint*, 2012.

[21] S. Gadat, F. Panloup, and C. Pellegrini. Large deviation principle for invariant distributions of memory gradient diffusions. *Preprint*, 2012.

Chapitres de livre

[22] J. Bigot and S. Gadat. *chapter : Pattern recognition through large deformations of images, in book Pattern Recognition.* Intech, 2010.

[23] S. Gadat. *chapter : Feature Selection in high dimension for face Detection, in book Advances in Face Image Analysis.* Techniques and Technologies, IGI - Global, 2009.

[24] 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.

Actes de conférences

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] S. Gadat. Markov hybrid process for variable selection in classification. In *Proceedings of the 47th Conference on Decision and Control*, Cancun, Mexico, December 2008.

[33] 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.

[34] 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.

[35] 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.

Rapports techniques

[36] 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.

[37] 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.

[38] 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.

Références

[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.