Classification supervisée et estimation non-paramétrique pour les EDS

Eddy Ella Mintsa

Université Gustave Eiffel

LAMA LABORATOIRE D'ANALYSE ET DE MATHÉMATIQUES APPLIQUÉES – UMR8050

École Doctorale Mathématiques et STIC

THÈSE

Présentée pour l'obtention du grade de DOCTEUR
de L'Université Gustave Eiffel

Spécialité

Mathématiques Appliquées

par

EDDY MICHEL ELLA MINTSA

Classification supervisée et estimation non-paramétrique pour les EDS

Soutenue le 12/12/2023 devant le jury composé de

ARNAK DALALYAN	PR	ENSAE	EXAMINATEUR
CHRISTOPHE DENIS	MCF HDR	UNIVERSITÉ GUSTAVE EIFFEL	CO-DIRECTEUR
CHARLOTTE DION-BLANC	MCF	SORBONNE UNIVERSITÉ	CO-DIRECTRICE
MARC HOFFMANN	PR	UNIVERSITÉ PARIS DAUPHINE	EXAMINATEUR
CLAIRE LACOUR	PR	UNIVERSITÉ GUSTAVE EIFFEL	EXAMINATRICE
DASHA LOUKIANOVA	MCF HDR	UNIVERSITÉ D'EVRY VAL D'ESSONNE	RAPPORTRICE
CLÉMENT MARTEAU	PR	UNIVERSITÉ LYON I	RAPPORTEUR
VIET-CHI TRAN	PR	UNIVERSITÉ GUSTAVE EIFFEL	DIRECTEUR

Remerciements

Le moment est venu de me rappeler, avec reconnaissance, de toutes les personnes qui m'ont soutenu de diverses manières, accompagné, instruit, encouragé, conseillé durant mon parcours à l'Université Gustave Eiffel, et particulièrement pendant ma thèse.

Tout d'abord, je tiens à exprimer ma profonde gratitude à mes encadrants de thèse, Christophe Denis, Charlotte Dion-Blanc et Viet-Chi Tran. J'ai eu la grâce de vous avoir comme co-directeurs de thèse, toujours disponibles en cas de besoin, y compris les week-ends, très attentionnés, toujours sereins et optimistes sur la bonne fin de cette thèse. Je n'oublierai pas la patience dont vous avez fait preuve, reconnaissant avoir très souvent montré des signes de paresse. J'ai beaucoup appris de votre expérience dans la recherche, et de votre immense connaissance en Statistiques et Probabilités. Je vous remercie de m'avoir proposé, après mon stage de recherche avec vous, un sujet de thèse sur une thématique à laquelle je porte un grand intérêt, et surtout d'avoir trouvé en moi, un candidat capable de porter un tel projet sous votre direction, et d'en obtenir des résultats satisfaisants. J'ai beaucoup appris de vos conseils, de vos orientations, et de vos réprimandes. Vous avez parfaitement joué votre rôle, et c'est une expérience que je ne suis pas prêt d'oublier. Je souhaite d'ailleurs continuer à travailler avec vous dans le futur.

Un grand merci à mes rapporteurs de thèse, Dasha Loukianova et Clément Marteau. Le travail immense que vous avez accompli est très cher à mes yeux et surtout très valorisant pour mon manuscrit de thèse. Merci pour votre amitié, et vos conseils qui ont contribué à l'amélioration du manuscrit. Merci à Arnak Dalalyan, Claire Lacour et Marc Hoffmann d'avoir accepté d'examiner mon travail. Merci à vous tous d'avoir accepté de faire partie de mon jury de thèse. Merci Marc pour ton invitation pour un exposé au Séminaire Parisien.

Mes sincères remerciement à l'ensemble des membres du laboratoire LAMA et de l'ensemble du personnel de l'Université Gustave Eiffel. Merci à Olivier Guédon et au Président de l'Université Gustave Eiffel, Gilles Roussel, pour leur soutien depuis le début de mon stage de recherche et pendant ma thèse. Merci aux membres du comité de suivi de thèse du laboratoire pour les échanges valorisants que vous avons eu, pour vos conseils et suggestions pour le bon déroulement de ma thèse et pour l'après-thèse. Merci à Miguel Martinez pour sa contribution à l'avancement de mes travaux de thèse, merci à Mohamed Hébiri pour ses encouragements et conseils, et merci à Mathieu Fradélizi pour son soutien. Mes sincères remerciements à Audrey Patout et Ketty Cimonard pour leur accompagnement et leur bienveillance pendant toute la durée de ma thèse. Merci à Robert Aymar par qui j'ai connu mes encadrants de thèse. Merci à tous les doctorants du laboratoire LAMA, merci pour votre amitié, pour les bons moments que nous avons passés ensemble depuis le début de ma thèse, et merci de la compréhension dont vous avez fait preuve pour mes absences récurrentes au restaurant à l'heure du déjeûner. Merci à Arafat, Benjamin, Ahmed, Josué, pour leurs partages d'expériences, leurs conseils, et leurs orientations. J'ai passé des moments inoubliables à vos cotés pendant ces trois dernières années et j'en suis très ravi.

Merci aux membres du laboratoire LPSM de Sorbonne Université. J'ai apprécié l'ambiance qui y

règne, en particulier lors des réunions et exposés du Groupe de Travail des Thésards (GTT), ou lors des pauses déjeûner. Merci en particulier à Pierre pour ses partages d'expériences.

Merci à Christèle Etchegaray pour sa contribution à la motivation pour mon sujet de thèse. Merci à Myriam Sidibe et Sylvie Cach pour leur disponibilité, leur bienveillance et leur réactivité. Un grand merci à Karine Marot et Anna Lu' Mercier pour leur soutien inestimable lors de procédures administratives depuis mes premiers jours au laboratoire LAMA.

Enfin, merci à mes parents pour leur soutien sur plusieurs plans, bien avant ma thèse, et pendant toute la durée de la thèse.

Productions scientifiques liées à la thèse

Articles Soumis

- Nonparametric estimation of the diffusion coefficient from SDE paths
- Nonparametric multiclass classification procedure of the plug-in type for SDE paths

Article en cours de rédaction

- Nonparametric multiclass classification procedure of the ERM type for SDE paths

Package R

- SDEclassif

Contents

Abstract

This thesis deals with statistical studies based on functional data modelled by stochastic differential equations whose solutions are called diffusion processes. More precisely, we bring contributions on the resolution of the problem of nonparametric estimation of coefficients of a diffusion process in short time, and on the construction of nonparametric procedures for multiclass classification of diffusion paths. The thesis is divided into three main parts.

In the first part of the thesis, we focus on the nonparametric estimation of the diffusion coefficient of a time-homogeneous stochastic differential equation with a finite time horizon. Two frameworks are considered. First, we assume that only a single discrete path is observed, and second, we assume that N diffusion paths, generated independently, are observed. In each framework, we propose projection estimators of the square of the diffusion coefficient respectively on a compact interval, and on the real line, by minimizing a least squares contrast function. We establish the consistency of estimators, and derive rates of convergence under mild assumption on the diffusion coefficient. This first part paves the way for the construction of an empirical classification procedure of the plug-in type, which is based on the nonparametric estimation of the coefficients of the diffusion processes.

The second part proposes a plug-in type classification procedure for diffusion paths. We consider a diffusion process X that is a strong solution of a time-homogeneous stochastic differential equation with a unknown drift coefficient depending on a label Y taking values in a finite set of $\mathbb{N}$, and with an unknown diffusion coefficient, common to all classes. We propose, from a learning sample containing N independent observations of the diffusion process, a plug-in classifier based on nonparametric estimators of coefficients of the diffusion process, and on the estimator of the discrete law of the label Y. The observation window is the compact interval $[0, 1]$. We prove the consistency of the plug-in classifier, and establish a rate of convergence of order $N^{-1/5}$ when the coefficients of the diffusion process are unknown and Lipschitz. We also derive some faster rates of convergence under stronger assumptions on the drift and the diffusion coefficient.

The last part focuses on the construction of a nonparametric procedure of Empirical Minimization Risk (ERM) type for multiclass classification of diffusion paths. The setting is the same as in the second part, and the goal is to build an empirical classifier by minimizing the empirical risk of classification defined from the learning sample. We prove the consistency of the ERM-type classifier, and derive a rate of convergence of order $N^{-\beta/(2\beta+1)}$ over a Hölder space of smoothness parameter $\beta \geq 1$. We then establish, in the binary classification setup, and over the same Hölder space, a faster rate of convergence of order $N^{-4\beta/3(2\beta+1)}$ in binary classification, thanks to a margin assumption imposed on the regression function associated to the model.

Key words: supervised classification, nonparametric estimation, diffusion paths, plug-in, empirical risk minimization.

Résumé

Cette thèse porte sur des études statistiques basées sur les données fonctionnelles modélisées par des équations différentielles stochastiques dont les solutions sont appelées processus de diffusion. Nous apportons des contributions au problème d'estimation non-paramétrique des coefficients d'un processus de diffusion en temps court, et à la question de construction des procédures non-paramétriques de classification multiclasses pour les trajectoires de diffusion en temps court. La thèse est divisée en trois principales parties.

Dans la première partie de la thèse, on se focalise sur l'estimation non-paramétrique du coefficient de diffusion d'une équation différentielle stochastique homogène en temps avec un horizon temporel fini. Deux cadres d'étude sont envisagés. Premièrement, on suppose qu'une seule trajectoire de diffusion est observée, et deuxièmement, on suppose observer N trajectoires de diffusion indépendantes. Dans chaque cadre d'étude, nous proposons des estimateurs par projection du carré du coefficient de diffusion respectivement sur un intervalle compact, et sur la droite réelle $\mathbb{R}$ par la minimisation d'une fonction de contraste des moindres carrés. Nous établissons la consistance des estimateurs et des vitesses de convergences sous des hypothèses de régularité sur le coefficient de diffusion. Cette première partie ouvre ainsi la voie à la construction d'une procédure de classification empirique de type plug-in, qui repose sur l'estimation non-paramétrique des coefficients des processus de diffusion.

La deuxième partie propose une procédure de classification de type *plug-in* des trajectoires de diffusion. Nous partons d'un modèle de diffusion dans lequel un processus de diffusion X est solution forte d'une équation différentielle stochastique homogène en temps dont le coefficient de dérive est supposé inconnu et dépend de la classe Y, une variable aléatoire discrète de loi inconnue, et dont le coefficient de diffusion, aussi supposé inconnu, est commun à toutes les classes. On propose, à partir d'un échantillon constitué de N observations indépendantes du processus de diffusion, un classifieur de type *plug-in* basé sur des estimateurs non-paramétriques des coefficients de dérive et de diffusion, et de l'estimateur de la loi discrète de l'étiquette Y. Nous prouvons la consistance du classifieur *plug-in* et établissons une vitesse de convergence de l'ordre de $N^{-1/5}$ lorsque les coefficients du processus sont des fonctions sont inconnues et lipschitziennes. Nous établissons ensuite des vitesses plus rapides sous des hypothèses plus fortes sur les coefficients de dérive et de diffusion.

La dernière partie traite de la construction d'une procédure non-paramétrique de classification multiclasses de type ERM pour les trajectoires de diffusion. Nous restons dans le même cadre d'étude que celui de la deuxième part, et l'objectif est de construire un classifieur empirique par la minimisation du risque empirique (ERM) de classification. Nous prouvons la consistance du classifieur de type ERM et établissons une vitesse de convergence de l'ordre de $N^{-\beta/(2\beta+1)}$ sur un espace de Hölder de paramètre de régularité $\beta \geq 1$. Nous établissons ensuite, sur le même espace de Hölder de régularité $\beta \geq 1$, une vitesse plus rapide de l'ordre de $N^{-4\beta/3(2\beta+1)}$ en classification binaire, grâce à une hypothèse de marge imposée sur la fonction de régression associée au modèle.

Mots-clé : classification supervisée, estimation non-paramétrique, trajectoires de diffusion, *plug-in*, minimisation du risque empirique.

Chapitre 1

Introduction

Sommaire

1.1 Préliminaires

Depuis quelques années, nous voyons apparaître de plus en plus fréquemment dans de nombreux domaines scientifiques, des données modélisées comme des réalisations de variables aléatoires fonctionnelles (par exemple des courbes ou des images). Ces données sont généralement le résultat de mesures répétées d'une quantité ou d'une grandeur quelconque en faisant varier un paramètre tel que le temps, une longueur (par exemple la longueur d'onde), etc. En théorie, les fonctions aléatoires qui génèrent ces données prennent leurs valeurs dans des espaces vectoriels de dimension infinie. Toutefois, dans le cas pratique, ces données sont représentées par des vecteurs de valeurs de très grande taille, chaque coordonnée étant le résultat d'une mesure en fonction du paramètre d'intérêt. Le progrès technologique a rendu possible une collecte rapide de ce type de données dans des domaines tels que la chimie, la météorologie, la biologie (see *e.g.* [39], [78], [79], [82]), ou encore la finance (voir par exemple [65]). Depuis plusieurs années, des données fonctionnelles font l'objet de nombreuses études statistiques telles que l'estimation paramétrique (voir par exemple [46], [60], [15], [51], [87]),

l'estimation non-paramétrique (voir par exemple [20], [42], [30], [88]), l'analyse discriminante ou la classification (voir par exemple [1], [86], [13], [72]).

Dans ce manuscrit de thèse, nous nous focalisons sur la classification multiclasses de données fonctionnelles générées par le modèle de mélange de diffusions suivant :

$$dX_t = b^*_Y(X_t)dt + \sigma^*(X_t)dW_t, \quad t \in [0,T], \quad X_0 = x_0 \tag{1.1}$$

où $(W_t)_{t\geq 0}$, appelé mouvement brownien standard, est un processus continu à accroissements indépendants et stationnaires défini sur un espace de probabilité filtré $(\Omega, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, où $(\mathcal{F}_t)_{t\geq 0}$ est la filtration naturelle de $(W_t)_{t\geq 0}$ (voir par exemple [66]). Ainsi, pour tout $s, t \geq 0$ tels que $t > s$, on a

$$W_t - W_s \perp \mathcal{F}_s, \quad \text{et} \quad W_t - W_s \overset{\mathcal{L}}{=} W_{t-s} \sim \mathcal{N}(0, t-s).$$

De plus, le mouvement brownien $(W_t)_{t\geq 0}$ est une martingale (voir [66]) sur sa filtration naturelle. La variable aléatoire Y est indépendante de $(W_t)_{t\geq 2}$ et à valeurs dans un ensemble fini $\mathcal{Y} = \{1, \ldots, K\}$, avec $K \geq 2$, et de loi inconnue $\mathfrak{p}^* = (\mathfrak{p}^*_1, \ldots, \mathfrak{p}^*_K)$, avec $\mathfrak{p}^*_0 = \min_{i\in\mathcal{Y}} \mathfrak{p}^*_i > 0$. Les fonctions de dérive $b^*_i,\ i \in \mathcal{Y}$ et le coefficient de diffusion σ^* sont supposés inconnus et appartiennent à un espace de fonctions de dimension infinie.

Les données de diffusion trouvent de nombreuses applications dans des domaines tels que la finance avec la modélisation de taux d'intérêt, l'étude du risque d'un actif financier, le coefficient de diffusion, encore appelé volatilité étant vu comme une mesure de risque de l'actif étudié (voir par exemple [65], [9]).

Dans ce manuscrit, nous proposons des procédures de classification non-paramétriques pour des trajectoires de diffusion essentiellement homogènes en temps, l'horizon temporel étant fixé à $T = 1$. Le terme non-paramétrique renvoie à un cadre dans lequel les coefficients de dérive et de diffusion des solutions d'équations différentielles stochastiques sont supposés inconnus, et ne dépendent pas d'un paramètre variant dans un espace de dimension finie. Nous montrons la consistance des classifieurs empiriques obtenus sous des hypothèses les plus faibles possibles sur les coefficients des processus de diffusion, et nous établissons des vitesses de convergence sous différentes hypothèses sur les coefficients des processus. La particularité de ces classifieurs empiriques est le fait qu'ils tirent profit des propriétés des processus de diffusion générant les observations à partir desquelles ils sont construits. Nous comparerons ainsi, au moyen d'études numériques sur des données simulées, la performance des classifieurs obtenus avec certaines méthodes de classification pour des données fonctionnelles trouvées dans la littérature scientifique.

Dans la section 1.2 nous parlons d'inférence statistique pour les équations différentielles stochastiques, et présentons particulièrement les modèles d'estimation non-paramétriques des coefficients de dérive et de diffusion. La section 1.3 est consacrée à la présentation de la classification supervisée dans un cadre général, et à la description du modèle de diffusion étudié dans cette thèse pour la classification de trajectoires de diffusions. Dans la section 1.4, nous présentons les principales contributions de la thèse, et enfin, une conclusion et les perspectives de recherche sont données en section 1.5. Les preuves de certains résultats sur les équations différentielles stochastiques sont données en annexes (section 1.6).

1.2 Inférence statistique pour les équations différentielles stochastiques

1.2.1 Modèle et hypothèse

On se réfère principalement à [81] et [66] pour l'étude des équations différentielles stochastiques. Comme nous l'avons dit plus haut, nous proposons dans ce manuscrit des procédures empiriques de classification supervisée pour des données fonctionnelles modélisées par des solutions d'équations différentielles stochastiques. Nous considérons plus particulièrement les diffusions homogènes en temps, solutions du modèle de mélange 1.1. On rappelle que le processus $\left(\int_0^t \sigma^*(X_s)dW_s\right)_{t\geq 0}$ est une martingale. Il en résulte que le processus de diffusion X solution de l'équation (1.1) est une semi-martingale et un processus de Markov (voir par exemple [81]), et ses coefficients b^*_Y et σ^* satisfont les égalités suivantes.

Proposition 1.2.1. *Supposons que $b_i^*,\ i \in \mathcal{Y}$ et σ^* sont des fonctions continues et fixons $s \in [0,1)$. On a alors*

$$\lim_{t \to s^+} \mathbb{E}\left[\frac{X_t - X_s}{t-s} - b_Y^*(X_s)\right] = 0, \tag{1.2}$$

$$\lim_{t \to s^+} \frac{\langle X, X\rangle_t - \langle X, X\rangle_s}{t-s} = \sigma^{*2}(X_s) \tag{1.3}$$

où pour tout $t \in [0,1]$, $\langle X, X\rangle_t = \int_0^t \sigma^{*2}(X_u)du$ est la variation quadratique à l'instant t du processus X.

L'équation (1.2) implique que pour tout instant t au voisinage de $s \in (0,1)$, on obtient, pour chaque étiquette $i \in \mathcal{Y}$ et conditionnellement à $\{Y = i\}$, le résultat approximatif ci-dessous :

$$\frac{X_t - X_s}{t-s} \approx b_i^*(X_s).$$

On voit ainsi que pour tout $s \in (0,1)$ et conditionnellement à Y, la quantité aléatoire $b_Y^*(X_s)$ peut être interprétée comme la dérivée à l'instant s du processus de diffusion $X = (X_t)_{t\in[0,1]}$. D'après l'équation (1.3) et l'expression de la variation quadratique en fonction de σ^{*2}, on déduit d'une part que la variation quadratique $t \mapsto \langle X, X\rangle_t$ est une fonction croissante du temps, et d'autre part, pour tout $t \in (0,1]$, $\langle X, X\rangle_t$ augmente avec le carré du coefficient de diffusion σ^{*2}. Ainsi, le coefficient de diffusion est vu comme une mesure des amplitudes des variations du processus de diffusion X lorsqu'on fait varier le temps t dans $[0,1]$. Plus précisément, une explosion des valeurs de $\sigma^{*2}(X_s)$ pour tout $s \in (0,1)$ entraine de fortes variations du processus de diffusion X lorsqu'on passe d'un instant $s \in (0,1)$ à un instant $t \in (s,1]$. Inversement, si les valeurs de $\sigma^{*2}(X_s)$ pour tout $s \in (0,1)$ sont proches de 0, alors on observe de très faibles variations des valeurs prises par $X_t - X_s$ pour tout $s, t \in [0,1]$ tels que $s < t$.

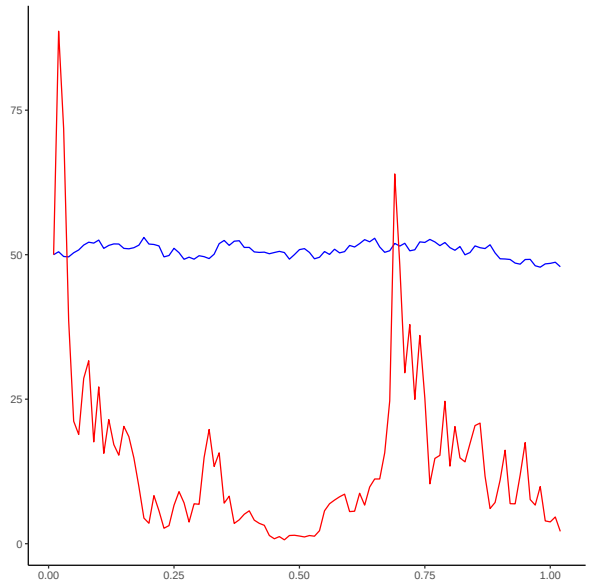

FIGURE 1.1 : La figure ci-dessus représente deux trajectoires de diffusion. Celle en bleu est une observation d'une solution de l'équation différentielle stochastique $dX_t = 0.15X_t dW_t$ avec $b_Y^* \equiv 0$ et $\sigma^* = 0.15$, et celle en rouge est générée par une solution de l'équation $dX_t = 5X_t dW_t$ avec $b_Y^* \equiv 0$ et $\sigma^* = 5$. Cette figure met ainsi en évidence l'impact du coefficient de diffusion sur les amplitudes des variations du processus de diffusion.

Une manière de classifier les trajectoires de diffusion serait de les distinguer suivant les valeurs des coefficients de dérive b_i^* ou des coefficients de diffusion σ^* des équations différentielles stochastiques qui les génèrent. Si nous prenons l'exemple de la finance, où le processus de diffusion X est à valeurs positives et modélise le prix d'un actif financier. Alors, la trajectoire en bleu de la figure 1.1 représentera les variations du prix d'un actif à faible risque avec une volatilité faible ($\sigma^* \equiv 0.15$),

tandis que la trajectoire en rouge représentera l'évolution du prix d'un actif à haut risque, avec une volatilité suffisamment élevée ($\sigma^* \equiv 5$). On peut ainsi, à partir de données observées, construire une procédure de classification qui distinguera les actifs à faible risque des actifs très risqués. Certaines de ces procédures de classification peuvent être basées sur des estimateurs de $b_i^*,\ i \in \mathcal{Y}$ ou σ^{*2}.

Supposons que les coefficients b_Y^* et σ^* du processus de diffusion solution de l'équation 1.1 satisfont l'hypothèse suivante.

Hypothèse 1.2.2. *(Ellipticité and regularité)*

(i) Il existe $L_0 > 0$ tel que les fonctions $b_i^*,\ i \in \mathcal{Y}$ et σ^* sont L_0-Lipschitziennes

$$\max_{i\in\mathcal{Y}} |b_i^*(x) - b_i^*(y)| + |\sigma^*(x) - \sigma^*(y)| \leq L_0|x-y|,\ \forall (x,y) \in \mathbb{R}^2.$$

(ii) Il existe des constantes σ_0^*, σ_1^* telles que

$$0 < \sigma_0^* \leq \sigma^*(x) \leq \sigma_1^*,\ \ \forall x \in \mathbb{R}.$$

Sous l'hypothèse 1.2.2, nous obtenons un certain nombre de résultats à partir de l'équation différentielle stochastique (1.1). Le premier résultat est l'unicité de la solution de l'équation (1.1) (voir par exemple [63]). Le deuxième résultat est l'existence d'une densité de transition $p_X : (t,x) \in [0,1]\times\mathbb{R} \mapsto p_X(t,x)$ du processus de diffusion X qu'on peut approximer par des densités gaussiennes. En effet, nous obtenons de [52] le résultat suivant.

Proposition 1.2.3. *Sous l'hypothèse 1.2.2, il existe des constantes $c > 1$, et $C > 1$ telles que pour tout $t \in (0,1]$ et pour tout $x \in \mathbb{R}$,*

$$\frac{1}{C\sqrt{t}} \exp\left(-c\frac{x^2}{t}\right) \leq p_X(t,x) \leq \frac{C}{\sqrt{t}} \exp\left(-\frac{x^2}{ct}\right).$$

La preuve de ce résultat est fournie dans [52]. L'approximation de la densité de transition p_X donnée par la proposition 1.2.3 est très utilisée pour l'obtention de résultats de consistance des classifeurs de type *plug-in* pour des données de diffusion. Nous y reviendrons plus en détails dans les chapitres 2 et 3 de ce manuscrit. Nous énonçons un autre résultat important qui découle de l'hypothèse 1.2.2 et qui est en rapport avec les moments du processus de diffusion X.

Proposition 1.2.4. *Sous l'hypothèse 1.2.2 et pour tout $q \geq 1$, il existe une constante $C_q \geq 1$ telle que*

$$\mathbb{E}_X\left[\sup_{t\in[0,1]} |X_t|^q\right] \leq C_q.$$

Nous pouvons ainsi déduire de la proposition 1.2.4 que sous l'hypothèse 1.2.2, le processus de diffusion X admet des moments finis de tout ordre. Ce résultat est essentiel pour la majoration des risques d'estimation des coefficients $b_i^*,\ i \in \mathcal{Y}$ et σ^{*2} de la diffusion X. Nous y reviendrons largement dans les deuxième et troisième chapitres de ce manuscrit.

L'estimation des coefficients d'un processus de diffusion est largement étudiée dans la littérature scientifique. On distingue le cas paramétrique (voir par exemple [46], [60], [15], [51], [87], [29]), et le cas non-paramétrique (voir par exemple [42], [22], [20], [56], [58]). Pour le cas non-paramétrique, on trouve généralement comme méthodes utilisées pour la construction des estimateurs des fonctions $b_i^*,\ i \in \mathcal{Y}$ et σ^{*2}, la méthode par noyaux (voir par exemple [41], [70]), et la méthode par regression (voir par exemple [22], [30], [58]).

Dans ce manuscrit, l'estimation des fonctions $b_i^*,\ i \in \mathcal{Y}$ et σ^{*2} est essentiellement non-paramétrique et basée sur la méthode de regression.

1.2.2 Estimation non-paramétrique des fonctions de dérive

Considérons le processus de diffusion $X = (X_t)_{t\in[0,1]}$ venant du modèle de diffusion (1.1) avec $b_Y^* = b^*$. La fonction de dérive b^* est inconnue, et le coefficient de diffusion peut être supposé connu ou inconnu, et fait partie du terme de bruit pour le modèle d'estimation de b^*.

Soit n un entier naturel suffisamment grand, et $\bar{X} = (X_{k\Delta_n})_{0\le k\le n}$ une observation en temps discret du processus de diffusion X solution de l'équation (1.1), avec $\Delta_n = 1/n$ le pas de discrétisation du temps. Soit $(Z_{k\Delta_n})_{0\le k\le n}$ la suite des incréments de X donnée par

$$Z_{k\Delta_n} := \frac{X_{(k+1)\Delta_n} - X_{k\Delta_n}}{\Delta_n}, \quad k = 0, \dots, n-1.$$

Ainsi, le coefficient de dérive b^* du processus de diffusion X satisfait le résultat ci-dessous.

Proposition 1.2.5. *Pour tout $k = 0, \dots, n-1$, on déduit de l'équation (1.1) que*

$$Z_{k\Delta_n} = b^*(X_{k\Delta_n}) + \xi_{k\Delta_n} \tag{1.4}$$

avec,

$$\xi_{k\Delta_n} := \frac{1}{\Delta_n}\int_{k\Delta_n}^{(k+1)\Delta_n} (b^*(X_s) - b^*(X_{k\Delta_n}))ds + \frac{1}{\Delta_n}\int_{k\Delta_n}^{(k+1)\Delta_n} \sigma^*(X_s)dW_s.$$

La fonction de dérive b^* est alors estimée comme une fonction de régression solution du modèle (1.4) avec $\xi_{k\Delta_n}$ le terme d'erreur. Notons que le modèle (1.4) est utilisé que l'on soit en temps court ($T < \infty$) ou en temps long ($T \to \infty$). Pour ce qui est des observations, et pour ce qui est du temps court, on verra qu'il n'est pas possible d'obtenir un estimateur convergent de b^* à partir d'une seule trajectoire de diffusion. Ainsi, nous supposons dans ce manuscrit avoir à notre disposition des observations discrètes répétées de X, c'est-à-dire des copies indépendantes $\bar{X}_1, \dots, \bar{X}_N$ du processus $\bar{X}$, avec N qui tends vers l'infini. En effet, dans le cas d'observations répétées du processus de diffusion, on peut construire un estimateur non-paramétrique convergent de b^* lorsque N tend vers l'infini.

Démonstration. D'après le modèle (1.1), on a

$$dX_t = b^*(X_t)dt + \sigma^*(X_t)dW_t$$

et pour tout $k \in [\![0, n-1]\!]$,

$$\begin{aligned} X_{(k+1)\Delta_n} - X_{k\Delta_n} &= \int_{k\Delta_n}^{(k+1)\Delta_n} b^*(X_s)ds + \int_{k\Delta_n}^{(k+1)\Delta_n} \sigma^*(X_s)dW_s \\ &= \Delta_n b(X_{k\Delta_n}) + \int_{k\Delta_n}^{(k+1)\Delta_n} (b^*(X_s) - b^*(X_{k\Delta}))ds + \int_{k\Delta_n}^{(k+1)\Delta_n} \sigma^*(X_s)dW_s. \end{aligned}$$

On obtient le résultat attendu en divisant chaque membre par Δ_n. □

L'estimation non-paramétrique de la fonction de dérive b^* par la méthode de régression est largement étudiée dans la littérature scientifique. Nous présentons ci-dessous deux des nombreuses références axées sur ce sujet.

[Comte, Genon-Catalot (2021)] ([20]). Dans cet article, le processus de diffusion X solution de (1.1) est observée en temps continu, avec un horizon temporel fini ($T < \infty$). Les auteurs proposent, à partir d'un échantillon $(X_1, \dots, X_N)$ constitué de N copies indépendantes de X, un estimateur par projection de b^* sur un espace d'approximation de dimension finie $m \ge 1$ engendré par une base orthonormée. La particularité de cet article est d'établir une vitesse de convergence de l'estimateur par projection de $b^*\mathbb{1}_A$ de l'ordre de $N^{-s/(s+1)}$ sur un espace de régularité $s \ge 1$, lorsque le support de la fonction de dérive b^* est la droite réelle $\mathbb{R}$, et A est un interval compact ou non-compact de $\mathbb{R}$. L'obtention d'un tel résultat nécessite toutefois une troncature de la dimension de l'espace d'approximation pour l'établissement d'un terme de variance convergeant vers 0 lorsque N tend vers l'infini.

i) **Construction de l'estimateur tronqué de b^*.**

L'estimateur par projection de b^* noté $\widehat{b}_m$ pour chaque $m \geq 1$, est donné par $\widehat{b}_m = \sum_{\ell=0}^{m-1} \widehat{\theta}_\ell \phi_\ell$, où $(\phi_0, \ldots, \phi_{m-1})$ est une base orthonormée engendrant l'espace d'approximation, et le vecteur de coordonnées $\widehat{\theta} = \left(\widehat{\theta}_0, \ldots, \widehat{\theta}_{m-1}\right)$ est donné par

$$\widehat{\theta} = \widehat{\Psi}_m^{-1} \widehat{Z}_m$$

où le vecteur $\widehat{Z}_m$ est défini par

$$\widehat{Z}_m = \left(\frac{1}{NT} \sum_{i=1}^{N} \int_0^T \phi_j(X_i(u)) dX_i(u) \right),$$

et la matrice aléatoire $\widehat{\Psi}_m$ est donnée par

$$\widehat{\Psi}_m = \left(\frac{1}{NT} \sum_{i=1}^{N} \int_0^T \phi_j(X_i(u)) \phi_\ell(X_i(u)) du \right)_{0 \leq j,\ell \leq m-1}.$$

La matrice $\widehat{\Psi}_m$ est un estimateur sans biais de la matrice de Gram Ψ_m de la base $(\phi_0, \ldots, \phi_{m-1})$ définie pour tout entier $m \geq 1$ par

$$\Psi_m = \mathbb{E}\left[\widehat{\Psi}_m\right] = \left(\mathbb{E}\left[\frac{1}{T} \int_0^T \phi_j(X_s) \phi_\ell(X_s) ds \right] \right)_{0 \leq j,\ell \leq m-1}.$$

La dimension m de l'espace d'approximation est ensuite tronquée comme suit :

$$L(m) \left(\left\| \Psi_m^{-1} \right\|_{\text{op}} \vee 1 \right) \leq \mathfrak{c}_T \frac{NT}{\log(NT)},$$

avec $\mathfrak{c}_T = \dfrac{1 - \log(2)}{8T}$, et

$$L(m) := \sup_{x \in A} \sum_{\ell=0}^{m-1} \phi_\ell^2(x) < +\infty. \tag{1.5}$$

Ainsi, pour tout $m \geq 1$, en remplaçant la matrice de Gram Ψ_m par son estimateur $\widehat{\Psi}_m$, l'estimateur tronqué $\widetilde{b}_m$ est défini à partir de $\widehat{b}_m$ par

$$\widetilde{b}_m = \widehat{b}_m \mathbb{1}_{\Lambda_m}$$

où

$$\Lambda_m = \left\{ L(m) \left(\left\| \widehat{\Psi}_m^{-1} \right\|_{\text{op}} \vee 1 \right) \leq \frac{\mathfrak{c}_T NT}{\log(NT)} \right\}.$$

Notons que la quantité $L(m)$ donnée par l'équation (1.5) est, en effet, définie pour tout $m \geq 1$ par

$$L(m) = \sup_{h \in \mathcal{S}_m,\ \|h\|=1} \sup_{x \in A} h^2(x)$$

et dépend ainsi uniquement du choix de l'espace d'approximation $\mathcal{S}_m$ et non du choix d'une base parmi les bases orthonormées.

ii) **Vitesse de convergence de l'estimateur tronqué de $b^* \mathbb{1}_A$.**

Une vitesse de convergence de l'ordre de $N^{-s/(s+1)}$ est établie sous l'hypothèse que la fonction de dérive $b^* \mathbb{1}_A$, restreinte à A, appartient à l'espace $W_{f_T}^s(A, R)$ défini par

$$W_{f_T}^s(A, R) = \left\{ h \in \mathbb{L}^2(A, f_T(x)dx),\ \forall \ell \geq 1,\ \left\| h - h_\ell^{f_T} \right\|_{f_T}^2 \leq R\ell^{-s} \right\}$$

avec

$$\left\|h - h_\ell^{f_T}\right\|_{f_T}^2 = \int_A (h - h_\ell^{f_T})^2(x) f_T(x) dx,$$

où la fonction f_T est une densité de probabilité, et pour toute fonction $h \in \mathbb{L}^2(A, f_T(x)dx)$ et pour chaque $\ell \geq 1$, $h_\ell^{f_T}$ est le projeté orthogonal de h sur un sous-espace de $\mathbb{L}^2(A, f_T(x)dx)$ de dimension ℓ. Cette vitesse est établie grâce à la troncature de la dimension, ayant prouvé que $\mathbb{P}(\Lambda_m^c) = \mathrm{O}(N^{-7})$.

Estimateur adaptatif de $b^*\mathbb{1}_A$. L'estimateur adaptatif $\widetilde{b}_{\widehat{m}}$ est obtenu en sélectionant à partir des données, une dimension optimal $\widehat{m}$ dans un ensemble $\widehat{\mathcal{M}}_N(\mathfrak{o}_T)$ défini par l'équation suivante

$$\widehat{\mathcal{M}}_N(\mathfrak{o}_T) = \left\{ m \in \{1, 2, \ldots, NT\},\ c_\phi^2 m \left(\left\| \widehat{\Psi}_m^{-1} \right\|_{\mathrm{op}} \vee 1 \right) \leq \mathfrak{o}_T \frac{NT}{\log(NT)} \right\}$$

où c_ϕ est la borne supérieure des fonctions de la base générant l'espace d'approximation, et la constante $\mathfrak{o}_T$ dépend de l'horizon temporel T et de la norme infinie $\|f_T\|_\infty$ de la fonction de densité f_T.

[Denis, Dion, Martinez (2020)] ([30]). Cet article propose, en temps court ($T < \infty$) et à partir de données répétées, un estimateur par projection de la fonction de dérive b^* sur un espace de dimension finie engendré par une base de fonctions **B**-spline et inclus dans l'espace initial contenant b^*. Une contrainte de type ℓ^2 est imposée aux vecteurs de coordonnées des éléments du sous-espace d'approximation, suivie d'une troncature de la borne de l'estimateur non-paramétrique obtenu, de sorte que celle-ci soit à croissance au plus polynômiale. On précise tout de même que la troncature effectuée sur l'estimateur de b^* n'a aucun impact sur la dimension de l'espace d'approximation. L'objectif de la contrainte ℓ^2 et de la troncature est d'assurer la convergence du risque d'estimation de l'estimateur non-paramétrique $\widehat{b}$ obtenu, et d'en établir, sous certaines hypothèses sur b^* et σ^*, une vitesse de convegence optimale de l'ordre de $N^{-\beta/(2\beta+1)}$ sur l'espace des fonctions de type Hölder de paramètre de régularité $\beta \geq 1$.

iii) **Construction de l'estimateur ridge de b^*.**

Partant d'un échantillon de trajectoires de diffusion $\left(\bar{X}^{(1)}, \ldots, \bar{X}^{(N)}\right)$ de taille N, observées de manière indépendante et en temps discret. Ils construisent, en utilisant le modèle de regression (1.4) et la minimisation de la fonction de constraste des moindres carrés obtenue à partir du modèle de régression, l'estimateur ridge $\widehat{b}_m = \sum_{\ell=0}^{m-1} \widehat{a}_\ell B_\ell$, où $(B_0, \ldots, B_{m-1})$ est la base de fonctions **B**-spline de l'espace d'approximation de dimension m, et

$$\widehat{\mathbf{a}} = (\widehat{a}_0, \ldots, \widehat{a}_{m-1}) = \underset{\|a\|_2^2 \leq mL}{\arg\min} \|\mathbf{Z} - \mathbf{B}\mathbf{a}\|_2^2$$

avec $\mathbf{Z}$ le vecteur des variables réponses donné par

$$\mathbf{Z} := \left(Z_{0\Delta_n}^{(1)}, \ldots, Z_{(n-1)\Delta_n}^{(1)}, \ldots, Z_{0\Delta_n}^{(N)}, \ldots, Z_{(n-1)\Delta_n}^{(N)} \right),$$

et la matrice $\mathbf{B}$ donnée ci-dessous

$$\mathbf{B} := \left({}^t\left(B_\ell(X_{0\Delta_n}^{(j)}), \ldots, B_\ell(X_{(n-1)\Delta_n}^{(j)}) \right) \right)_{\substack{1 \leq j \leq N \\ 0 \leq \ell \leq m-1}}.$$

La constante $L > 0$ et la dimension m de l'espace d'approximation sont des suites croissantes de la taille N de l'échantillon des observations. Enfin, l'estimateur obtenu est indépendant du support de la fonction de dérive b^*.

ii) **Consistance de l'estimateur ridge de b^*.**

La consistance de l'estimateur ridge $\widehat{b}_m$ est établie lorsque la fonction de dérive b^* a pour support la droite réelle $\mathbb{R}$. Ce résultat de consistance est obtenu entre autre, sous des hypothèses sur la dimension m de l'espace d'approximation, la constante L, et l'intervalle borné sur lequel est définie la base de fonctions **B**-splines qui engendre l'espace d'approximation.

iii) **Vitesse de convergence de l'estimateur ridge de b^*.**

L'établissement de la vitesse optimale de l'estimateur $\widehat{b}_m$ nécessite, sous des hypothèses requises sur b^* et σ^*, une approximation de la densité de transition du processus de diffusion X solution de (1.1) par des gaussiennes, une restriction de la fonction de dérive b^* à un intervalle compact comme par exemple $[0,1]$, et une définition appropriée d'un ensemble $\mathcal{M}$ des valeurs possibles de la dimension m de l'espace d'approximation. Ils établissent ainsi, pour tout $m \in \mathcal{M}$, une vitesse optimale de l'estimateur $\widehat{b}_m$ de l'ordre de $N^{-\beta/(2\beta+1)}$ sur un espace Hölder de paramètre de régularité $\beta \geq 1$.

iv) **Estimateur adaptatif de b^*.**

Un estimateur adaptatif de b^* est ensuite proposé par une sélection, à partir des données observées, de la dimension de l'espace d'approximation. Cette sélection de modèle se fait par minimisation de la somme de la fonction de contraste des moindres carrés et d'une fonction de pénalité dont l'ordre est celui du terme de variance du risque d'estimation de $\widehat{b}_m$.

Dans ce manuscrit, et pour la construction d'un classifieur de type *plug-in*, nous nous inspirons principalement de ces deux articles pour construire des estimateurs non-paramétriques des coefficients de dérive du modèle de diffusion (1.1) dont le support est la droite réelle $\mathbb{R}$. Nous établissons la vitesse optimale obtenue dans [30] en utilisant un raisonnement similaire à celui décrit dans [20].

1.2.3 Estimation non-paramétrique du coefficient de diffusion

On considère de nouveau le processus de diffusion $X = (X_t)_{t\in[0,1]}$ solution du modèle (1.1) avec $b_Y^* = b^*$.

Soit donc n un entier naturel suffisamment grand et $\bar{X} = (X_{k\Delta_n})_{0\leq k\leq n}$ une observation discrète de X avec $\Delta_n = 1/n$ le pas de temps. Soit $(U_{k\Delta_n})_{0\leq k\leq n}$ la suite construite à partir de $\bar{X}$ telle que

$$U_{k\Delta_n} := \frac{(X_{(k+1)\Delta_n} - X_{k\Delta_n})^2}{\Delta_n}, \quad k = 0, \ldots, n-1.$$

Ainsi, pour tout $k \in [\![0, n]\!]$, $U_{k\Delta_n}$ est une approximation en temps discret de la variance quadratique instantanée $d\langle X, X\rangle_t / dt$. La proposition ci-dessous donne le modèle de régression à partir duquel, le carré du coefficient de diffusion σ^{*2} peut être estimé comme solution du modèle.

Proposition 1.2.6. *Supposons que la fonction σ^* est de classe $\mathcal{C}^2$ sur $\mathbb{R}$. Pour tout $k = 0, \ldots, n-1$, on déduit de l'équation (1.1) que*

$$U_{k\Delta_n} = \sigma^{*2}(X_{k\Delta_n}) + \zeta_{k\Delta_n} + R_{k\Delta_n} \tag{1.6}$$

où $\zeta_{k\Delta_n} = \zeta^{(1)}_{k\Delta_n} + \zeta^{(2)}_{k\Delta_n} + \zeta^{(3)}_{k\Delta_n}$ avec,

$$\zeta^{(1)}_{k\Delta_n} = \frac{1}{\Delta_n}\left[\left(\int_{k\Delta_n}^{(k+1)\Delta_n} \sigma^*(X_s)dW_s\right)^2 - \int_{k\Delta_n}^{(k+1)\Delta_n} \sigma^{*2}(X_s)ds\right]$$

$$\zeta^{(2)}_{k\Delta_n} = \frac{2}{\Delta_n}\int_{k\Delta_n}^{(k+1)\Delta_n} ((k+1)\Delta_n - s)\sigma^{*\prime}(X_s)\sigma^{*2}(X_s)dW_s$$

$$\zeta^{(3)}_{k\Delta_n} = 2b^*(X_{k\Delta_n})\int_{k\Delta_n}^{(k+1)\Delta_n} \sigma^*(X_s)\, dW_s,$$

et $R_{k\Delta_n} = R^{(1)}_{k\Delta_n} + R^{(2)}_{k\Delta_n}$ *avec,*

$$R^{(1)}_{k\Delta_n} = \frac{1}{\Delta_n}\left(\int_{k\Delta_n}^{(k+1)\Delta_n} b^*(X_s)ds\right)^2 + \frac{1}{\Delta_n}\int_{k\Delta_n}^{(k+1)\Delta_n}((k+1)\Delta_n - s)\Phi(X_s)ds$$
$$R^{(2)}_{k\Delta_n} = \frac{2}{\Delta_n}\left(\int_{k\Delta_n}^{(k+1)\Delta_n}(b^*(X_s) - b^*(X_{k\Delta_n}))\,ds\right)\left(\int_{k\Delta_n}^{(k+1)\Delta_n}\sigma^*(X_s)dW_s\right)$$

et $\Phi := 2b^*\sigma^{*\prime}\sigma^* + \left[\sigma^{*\prime\prime}\sigma^* + (\sigma^{*\prime})^2\right]\sigma^{*2}$.

Le modèle de régression (1.6) est utilisé en temps court ($T \in \mathbb{R}_+$) comme en temps long ($T \to \infty$) (voir [22]). Notons que le terme $\zeta^{(1)}_{k\Delta_n}$ est le véritable terme d'erreur dépendant à la fois de N et n, les autres termes $\zeta^{(2)}_{k\Delta_n}$ et $\zeta^{(3)}_{k\Delta_n}$ étant des résidus négligeables (voir par exemple [22]). Le terme d'erreur $R_{k\Delta_n}$ est directement lié au pas de discrétisation du temps Δ_n. Constrairement à la fonction de dérive, un estimateur consistant de σ^{*2} peut être construit à partir d'une seule observation discrète du processus de diffusion X (voir par exemple [58]). Nous proposons dans ce manuscrit, des estimateurs ridge de σ^{*2} dans le cas d'observation d'une seule trajectoire ($N = 1$ et $n \to \infty$), et dans le cas d'observations répétées du processus X ($n, N \to \infty$).

L'estimation du carré du coefficient de diffusion est étudiée dans la littérature scientifique tant pour le temps long ($T \to \infty$) que pour le temps court ($T \in \mathbb{R}_{+*}$). Nous présentons ci-dessous, deux articles dans lesquels sont proposés des estimateurs de σ^{*2} en utilisant un modèle de régression non-paramétrique.

Hoffmann (1999) [58]. L'auteur propose un estimateur non-paramétrique du carré du coefficient de diffusion σ^* lorsque la fonction de dérive dépend à la fois du temps et de l'espace. σ^{*2} est estimé sur l'intervalle compact $D = [0, 1]$, à partir d'une observation discrète $X^{(n)} = (X_{i/n})_{0\le i\le n}$ du processus de diffusion X. Pour tout $p \in [1, \infty)$, une vitesse de convergence optimale de l'ordre de $n^{-ps/2(1+2s)}$ est établie sur l'espace de Besov $\mathbf{B}_{sp\infty}(D)$ sur l'intervalle D de paramètre de régularité $s > 1 + 1/2$ et défini à partir de l'espace $\mathbb{L}^p(D)$. L'établissement de cette vitesse optimale repose sur l'utilisation de la base d'ondelettes qui génère l'espace de Besov, et est optimale au sens minimax.

i) **Construction de l'estimateur de σ^{*2} sur l'intervalle $D = [0, 1]$.**

Pour rappel, le processus de diffusion $X = (X_t)_{t\in[0,1]}$ prend ses valeurs dans la droite réelle $\mathbb{R}$. Ainsi, la première étape consiste à assurer une répartition uniforme des points de données dans l'intervalle $D = [0, 1]$. Cette condition étant vérifiée dans le cas général où les observations sont supposées admettre une densité minorée sur l'intervalle d'estimation, l'auteur considère le temps local du processus X donné par

$$\forall\, x \in D,\quad L^x = \lim_{\varepsilon\to 0}\frac{1}{2\varepsilon}\int_0^1 \mathbb{1}_{|X_s - x|\le\varepsilon}ds. \tag{1.7}$$

De plus, pour un réel fixé $\nu \in (0, 1)$, L'estimateur de σ^{*2} est construit conditionnellement à l'événement aléatoire $\{L^D \ge \nu\}$ où $L^D = \inf_{x\in D} L^x$.

Un seuil $h_n > 0$ est choisi, et l'intervalle D est divisé en $\lfloor h_n^{-1}\rfloor$ boites $C_\lambda,\ \lambda = 1, \ldots, \lfloor h_n^{-1}\rfloor$ de taille h_n. Le temps local empirique $x \mapsto L^x_n$ donné pour tout $x \in D$ par

$$L^x_n = \frac{1}{nh_n}\sum_{i=0}^{n}\mathbb{1}_{|X_{i/n} - x|\le h_n/2}$$

satisfait $L^x_n \to L^x$ pour tout $x \in D$ lorsque $h_n \to 0$ et $nh_n \to \infty$. Ainsi, sur l'événement $\{L^D \ge \nu\}$, au moins $\lfloor nh_n\nu\rfloor$ points de données se trouvent dans chaque boîte C_λ avec une probabilité élevée.

L'estimateur non-paramétrique de σ^{*2} sur l'intervalle $D = [0, 1]$ est

$$\widehat{\sigma}_n^2(x) = \sum_{k=0}^{N_0-1} \widehat{\alpha}_{jk}^{\ell}\phi_{jk}^{\ell}(x) + \sum_{k\in s_j} \widehat{\alpha}_{jk}\phi_{jk}(x) + \sum_{k=0}^{N_0-1} \widehat{\alpha}_{jk}^{r}\phi_{jk}^{r}(x) \tag{1.8}$$

avec

$$\widehat{\alpha}_{jk} = \sum_i \left(X_{(i+1)/n} - X_{i/n}\right)^2 \phi_{jk}\left(\frac{i}{n}\right),$$

où ϕ_{jk}^{ℓ}, ϕ_{jk} et ϕ_{jk}^{r} sont des ondelettes appartenant à la base du sous-espace V_j de l'espace de Besov $\mathbf{B}_{sp\infty}(D)$ de dimension 2^j avec $j \in \mathbb{N}^*$. Les fonctions ϕ_{jk}^{ℓ} et ϕ_{jk}^{r} sont respectivement pour chaque k, les fonctions de mise à l'échelle des bords gauche et droit de l'intervalle $D = [0, 1]$. Enfin, s_j est l'ensemble des k tels que $\text{supp}\phi_{jk} \subset D$.

ii) **Vitesse optimale de convergence de l'estimateur de σ^{*2} sur l'intervalle $D = [0, 1]$.**
La vitesse optimale de convergence de l'estimateur non-paramétrique $\widehat{\sigma}_n^2$ donné par l'équation (1.8) est établie sous certaines hypothèses de régularité sur les coefficients b^* et σ^* du processus de diffusion X, et en tirant profit de l'optimalité au sens minimax de la base d'ondelettes. Ainsi, l'auteur montre que l'estimateur $\widehat{\sigma}_n^2$ atteint une vitesse optimale de l'ordre de $n^{-ps/2(1+2s)}$ sur un espace de Besov $\mathbf{B}_{sp\infty}(D)$ de paramètre de régularité $s > 2$ basé sur l'espace $\mathbf{L}^p(D)$.

Nous étudions dans le chapitre 2 l'estimation non-paramétrique du coefficient de diffusion issu du modèle (1.1) à partir d'une seule trajectoire de diffusion. Cette étude est basée sur le temps local (1.7) dont on établit le lien avec la densité de transition du processus solution du modèle (1.1).

Comte, Genon-Catalot, Rozenholc (2007) [22]. Dans cet article, le processus X est supposé ergodique et strictement stationnaire, avec l'horizon temporel $T \to \infty$. En effet, les auteurs disposent d'une observation discrète $(X_{k\Delta_n})_{1\leq k\leq n+1}$ du processus X, où le pas de temps Δ_n satisfait

$$\Delta_n \to 0 \ \text{ et } \ n\Delta_n \to \infty \ \text{ lorsque } \ n \to \infty.$$

Ils proposent dans cet article, un estimateur par projection de la restriction $\sigma_A^{*2} = \sigma^{*2}\mathbb{1}_A$ du carré du coefficient de diffusion σ^{*2}, solution du modèle de régression (1.6), par la minimisation d'une fonction de contraste des moindres carrés, où $A = [0, 1]$. Plus précisément, l'estimateur par projection $\widehat{\sigma}_m^2$ de σ_A^{*2} sur un sous-espace vectoriel $\mathcal{S}_m$ de dimension finie $m \geq 1$ est construit tel que

$$\widehat{\sigma}_m^2 = \underset{h\in\mathcal{S}_m}{\arg\min}\widetilde{\gamma}_n(h), \ \text{ avec } \ \widetilde{\gamma}_n(h) = \frac{1}{n}\sum_{k=1}^{n}[U_{k\Delta_n} - h(X_{k\Delta_n})]^2.$$

Ils établissent ensuite, sous l'hypothèse que σ_A^{*2} appartient à un espace de Besov $\mathcal{B}_{\alpha,2,\infty}([0, 1])$ avec $\|\sigma_A^{*2}\|_{\alpha,2,\infty} < \infty$, de régularité $\alpha > 1/2$, le résultat de consistance suivant :

$$\mathbb{E}\left[\|\widehat{\sigma}_m^2 - \sigma_A^{*2}\|_n^2\right] = \mathrm{O}\left(n^{-2\alpha/(2\alpha+1)} + \Delta_n^2 + \frac{1}{n}\right),$$

où,

$$\|\widehat{\sigma}_m^2 - \sigma_A^{*2}\|_n^2 = \frac{1}{n}\sum_{k=1}^{n}(\widehat{\sigma}_m^2 - \sigma_A^{*2})^2(X_{k\Delta_n}).$$

Dans cette thèse, on se place dans un cadre d'étude plus proche de celui de Hoffmann [58]. Toutefois, les méthodes de construction des estimateurs ainsi que les méthodes d'établissement de résultats théoriques sont similaires à celles utilisées dans dans Comte, Genon-Catalot et Rozenholc [22]. La section suivante présente quelques généralités sur la classification supervisée et multiclasses.

1.3 Généralités sur la classification supervisée

1.3.1 Principe

On entend par classification, un tri d'objets généralement appelés individus (ou feature en anglais) selon un critère de similarité défini au préalable. On appelle donnée étiquetée ou encore donnée labellisée, un couple de variables aléatoires $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ défini sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, où la variable $X \in \mathcal{X}$ est appelée feature ou individu, et la variable aléatoire $Y \in \mathcal{Y}$ est discrète, et est appelée étiquette ou classe. L'ensemble $\mathcal{Y}$ des classes est donnée par $\mathcal{Y} = \{1, \dots, K\}$ avec $K \geq 2$ le nombre total de classes, et on parle de classification multiclasses lorsque K est au moins égal à 3, et de classification binaire lorsque $K = 2$. Une classification est dite supervisée lorsqu'elle est basée sur des données étiquetées. Ce mode de classification est à différencier de la classification non-supervisée qui sert généralement à organiser des individus en groupes homogènes, les classes n'étant pas connues d'avance, et sont alors définies dans la procédure de classification. Elle a souvent pour finalité l'étude d'une population et de sa structure, et est basée sur des données non étiquetées.

La loi du couple aléatoire (X, Y) est supposée inconnue. Ainsi, nous supposons avoir à notre disposition un échantillon $\mathcal{D}_N = \big((X^1, Y_1), \dots, (X^N, Y_N)\big)$ de taille N et appelé échantillon d'apprentissage, et à partir duquel on entraîne puis teste une procédure empirique de classification pour la prédiction des étiquettes des nouvelles données.

La figure 1.2 présente un exemple d'échantillon d'apprentissage lorsque la variable $X = (X_t)_{t\in[0,1]}$ est un processus de diffusion à valeurs dans $\mathbb{R}$, solution d'une équation différentielle stochastique, et l'étiquette Y est une variable aléatoire distribuée de manière équiprobable dans $\mathcal{Y} = \{1, 2, 3\}$. Ainsi, on entraine le classifieur empirique $\widehat{g}$ à partir de l'échantillon qu'on observe à gauche de la figure, pour ensuite prédire à l'aide de $\widehat{g}$ une classe pour la nouvelle trajectoire de diffusion présentée à droite.

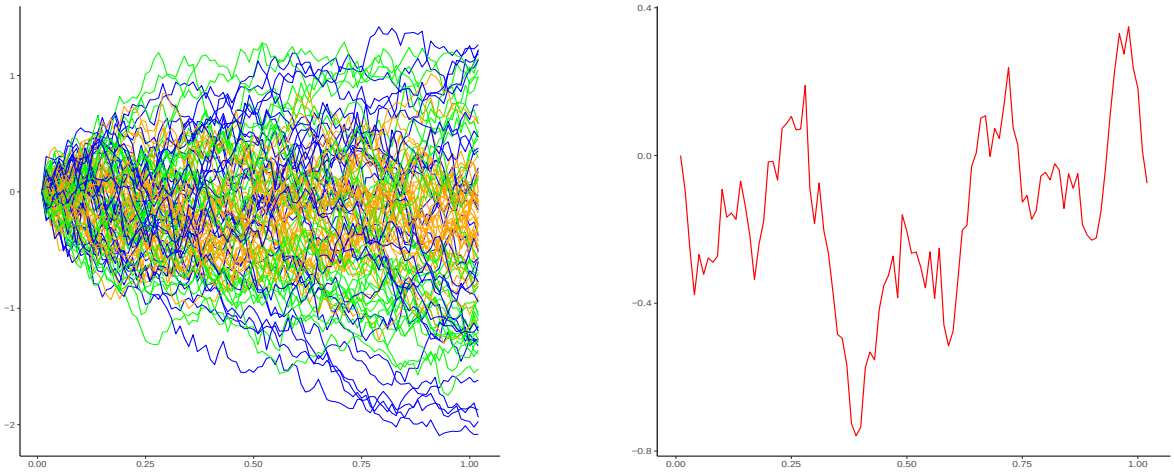

FIGURE 1.2 : *Du coté gauche, un échantillon d'apprentissage constitué de trajectoires de diffusions regroupées en trois classes. Du coté droit, une nouvelle trajectoire de diffusion dont la classe doit être prédite.*

On appelle classifieur ou règle de classification, toute fonction g mesurable sur l'espace des individus $\mathcal{X}$ et à valeurs dans l'ensemble des classes $\mathcal{Y}$. Ainsi, tout classifieur g prend en argument un individu X et lui alloue une classe $g(X)$ dans $\mathcal{Y}$. On note $\mathcal{G}(\mathcal{X}, \mathcal{Y})$, l'ensemble des classifieurs mesurables sur $\mathcal{X}$ et à valeurs dans $\mathcal{Y}$. La prédiction d'une classe par un classifieur g admet un coût évaluée par une fonction de perte $\ell : (y, y') \in \mathcal{Y} \times \mathcal{Y} \mapsto \ell(y, y') \in \mathbb{R}_+$. Ainsi, la performance de tout classifieur $g \in \mathcal{G}(\mathcal{X}, \mathcal{Y})$ est mesurée par une fonction $\mathcal{R} : \mathcal{G}(\mathcal{X}, \mathcal{Y}) \longrightarrow \mathbb{R}_+$ qui, à toute règle de classification $g \in \mathcal{G}(\mathcal{X}, \mathcal{Y})$ associe la quantité

$$\mathcal{R}(g) = \mathbb{E}\left[\ell(g(X), Y)\right].$$

La fonction $\mathcal{R}$ est appelée risque de classification, ou erreur de mauvaise classification, et pour tout classifieur $g \in \mathcal{G}(\mathcal{X}, \mathcal{Y})$, $\mathcal{R}(g)$ est par définition son coût moyen de prédiction. Dans cette thèse, nous

considérons la fonction de coût $\ell : (y, y') \in \mathcal{Y} \mapsto \mathbb{1}_{y \neq y'}$. Ainsi, le risque de classification $\mathcal{R}$ est donné pour chaque $g \in \mathcal{G}(\mathcal{X}, \mathcal{Y})$ par

$$\mathcal{R}(g) = \mathbb{P}(g(X) \neq Y). \tag{1.9}$$

1.3.2 Classifieur de Bayes et premières propriétés

Nous présentons dans cette section le classifieur de Bayes, et la propriété centrale qui met en relation un classifieur quelconque et le classifieur de Bayes. L'objectif de cette relation est l'étude de la performance du classifieur quelconque en se référant au risque du classifieur de Bayes. Nous renvoyons le lecteur à [32] pour une étude sur le classifieur de Bayes.

Definition 1.3.1 (Classifieur de Bayes)**.**
Supposons que la fonction $g \mapsto \mathcal{R}(g)$ atteint un minimum sur l'ensemble $\mathcal{G}(\mathcal{X}, \mathcal{Y})$ des classifieurs $g : \mathcal{X} \to \mathcal{Y}$. On appelle classifieur de Bayes, un classifieur $g^* \in \mathcal{G}(\mathcal{X}, \mathcal{Y})$ qui minimise le risque de classification $\mathcal{R}$ donné par l'équation (1.9).

Le classifieur de Bayes g^* est donc un classifieur optimal avec lequel nous pouvons classifier les données en prenant le moins de risque possible (voir par exemple [32]). Ce classifieur satisfait la relation équivalente suivante :

$$g^* \in \underset{g \in \mathcal{G}(\mathcal{X}, \mathcal{Y})}{\arg\min}\ \mathcal{R}(g).$$

Remarquons, pour tout classifieur $g \in \mathcal{G}(\mathcal{X}, \mathcal{Y})$, que

$$\mathcal{R}(g) = \mathbb{E}\left[\mathbb{1}_{Y \neq g(X)}\right] = \sum_{i=1}^{K} \mathbb{E}\left[\mathbb{1}_{g(X) \neq i} \mathbb{P}(Y = i|X)\right] = 1 - \sum_{i=1}^{K} \mathbb{E}\left[\mathbb{1}_{g(X) = i} \mathbb{P}(Y = i|X)\right]. \tag{1.10}$$

Ainsi, le classifieur de Bayes prend comme argument un individu $X \in \mathcal{X}$, évalue la probabilité d'appartenance de X à chacune des classes $i \in \mathcal{Y}$, et renvoie la classe $g^*(X) \in \mathcal{Y}$ pour laquelle la probabilité d'appartenance est maximale. Formellement, g^* est caractérisé pour tout $X \in \mathcal{X}$ par

$$g^*(X) = \underset{i \in \mathcal{Y}}{\arg\max}\ \mathbb{P}(Y = i|X). \tag{1.11}$$

Pour tout $i \in \mathcal{Y}$, on note,

$$\pi_i^*(X) := \mathbb{P}(Y = i|X). \tag{1.12}$$

Le résultat suivant donne la formule de l'excès de risque de tout classifieur $g \in \mathcal{G}(\mathcal{X}, \mathcal{Y})$ par rapport au classifieur de Bayes g^*.

Proposition 1.3.2. *Pour tout classifieur $g \in \mathcal{G}(\mathcal{X}, \mathcal{Y})$, on a :*

$$\mathcal{R}(g) - \mathcal{R}(g^*) = \mathbb{E}\left[\sum_{i=1}^{K} \sum_{k \neq i} |\pi_i^*(X) - \pi_k^*(X)|\, \mathbb{1}_{g^*(X) = i} \mathbb{1}_{g(X) = k}\right].$$

La preuve de la proposition 1.3.2 est faite dans [29] pour le cas des données de diffusion avec $X = (X_t)_{t \in [0,T]}$ et $T > 0$ l'horizon temporel, et elle reste exactement la même dans le cas général.

L'excès de risque donné par la proposition 1.3.2 est particulièrement utilisé pour l'étude théorique de la performance d'une règle de classification empirique construite à partir d'un échantillon de données.

1.3.3 Classifieur empirique

La loi du couple aléatoire (X, Y) est supposée inconnue. On part ainsi d'un échantillon d'apprentissage $\mathcal{D}_N$ pour entrainer une règle de classification empirique $\widehat{g}$ qui servira à faire des prédictions de classes pour les nouvelles observations du couple (X, Y). La règle de classification empirique est construite de manière à imiter le classifieur de Bayes, de sorte que son risque de classification $\mathcal{R}(\widehat{g})$ converge, selon la norme L^1, vers le risque du classifieur de Bayes $\mathcal{R}(g^*)$. Plus précisément, on construit à partir de l'échantillon $\mathcal{D}_N$, les estimateurs des probabilités conditionnelles $\pi_i^*(X)$, et l'estimateur $\widehat{g}$ est caractérisé pour toute observation $X \in \mathcal{X}$ par

$$\widehat{g}(X) = \arg\max_{i \in \mathcal{Y}} \widehat{\pi}_i(X).$$

On évalue théoriquement la performance de $\widehat{g}$ en regardant son excès de risque $\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)$ par rapport au classifieur de Bayes.

Ainsi, la règle de classification $\widehat{g}$ n'aura d'intérêt que si son excès de risque $\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)$ tend vers zéro lorsque la taille N de l'échantillon $\mathcal{D}_N$ tend vers l'infini. Des études supplémentaires sont réalisées pour établir la vitesse de convergence du classifieur $\widehat{g}$ sous certaines hypothèses imposées au modèle étudié.

1.3.4 Méthodes de construction de procédures de classification

Nous présentons dans cette section deux méthodes de construction de procédures de classification empiriques étudiées dans cette thèse. Le but de ces méthodes est de construire, à partir d'un échantillon d'apprentissage, un classifieur empirique qui imite le classifieur de Bayes. Le problème qui se pose est que la loi $\mathbb{P}$ des observations est inconnue.

Classifieurs de type *plug-in*. On suppose que la loi des observations $\mathbb{P} = \mathbb{P}_{\theta^*}$ dépend d'un paramètre inconnu θ^* de dimension finie ou infinie. Ainsi, on note le classifieur de Bayes $g^* = g_{\theta^*}$ et les probabilités $\pi_i^*(X) = \pi_{\theta^*}(i)$, et pour tout $X \in \mathcal{X}$,

$$g_{\theta^*}(X) = \arg\max_{i \in \mathcal{Y}} \pi_{\theta^*}(i).$$

On appelle classifieur *plug-in*, le classifieur $\widehat{g} = g_{\widehat{\theta}}$ obtenu à partir du classifieur de Bayes en remplaçant le paramètre inconnu θ^* par son estimateur $\widehat{\theta}$ construit à partir des données observées. En conséquence, la performance du classifieur empirique $\widehat{g}$, mesurée par le risque de classification $\mathcal{R} : g \mapsto \mathbb{P}_{\theta^*}(Y \neq g(X))$, dépendra fortement de la qualité d'estimation de θ^*.

Dans ce manuscrit, nous proposons une procédure de classification de type *plug-in*, où X est solution d'une équation différentielle stochastique dont les coefficients de dérive $b_i^*,\ i \in \mathcal{Y}$ et le coefficient de diffusion σ^* sont supposés inconnus, ainsi que la loi $\mathfrak{p}^*$ de l'étiquette $Y \in \mathcal{Y}$. Dans ce cas, le paramètre θ^*, de dimension infinie, est donné par

$$\theta^* = (\mathbf{b}^*, \sigma^{*2}, \mathfrak{p}^*),$$

avec $\mathbf{b}^* = (b_1^*, \ldots, b_K^*)$.

Classifieur de type ERM. La loi $\mathbb{P}$ des observations étant supposée inconnue, le risque de classification n'est donc pas calculable en pratique. Ainsi, partant d'un échantillon d'apprentissage $\mathcal{D}_N$ constitué de N copies indépendantes du couple aléatoire (X, Y), nous considérons le risque empirique de classification $\widehat{\mathcal{R}}_N : \mathcal{G}(\mathcal{X}, \mathcal{Y}) \longrightarrow \mathbb{R}_+$ donné pour tout classifieur g par

$$\widehat{\mathcal{R}}_N(g) := \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}_{Y_j \neq g(X_j)}.$$

Le risque empirique de classification $\widehat{\mathcal{R}}_N(g)$ d'un classifieur g est un estimateur sans biais du risque de classification $\mathcal{R}(g)$, et on a

$$\mathbb{E}\left[\left|\widehat{\mathcal{R}}_N(g) - \mathcal{R}(g)\right|\right] = \mathrm{O}\left(\frac{1}{\sqrt{N}}\right).$$

On remarque ainsi que pour tout $g \in \mathcal{G}(\mathcal{X}, \mathcal{Y})$, le risque empirique $\widehat{\mathcal{R}}_N(g)$ converge vers le risque théorique $\mathcal{R}(g)$ avec une vitesse de convergence considérée comme assez rapide et de l'ordre de $1/\sqrt{N}$. L'avantage ici, est que le risque empirique de classification $\widehat{\mathcal{R}}_N$ est calculable contrairement au risque théorique $\mathcal{R}$.

On appelle classifieur de type ERM, le classifieur empirique $\widehat{g}$ construit à partir de l'échantillon d'apprentissage $((X_1, Y_1), \ldots, (X_N, Y_N))$ tel que

$$\widehat{g} = \underset{g \in \mathcal{G}(\mathcal{X}, \mathcal{Y})}{\arg\min}\ \widehat{\mathcal{R}}_N(g). \tag{1.13}$$

Le classifieur de type ERM, reposant sur la minimisation du risque empirique de classification $\widehat{\mathcal{R}}_N$ est connu pour être plus performant que le classifieur de type *plug-in* décrit précédemment. Cependant, le problème d'optimisation (1.13) est non-convexe et donc très souvent incalculable en pratique. Ainsi, la construction d'une procédure de classification de type ERM implémentable passe par la convexification du problème d'optimisation (1.13) en utilisant une fonction de perte quadratique et en remplaçant l'espace des classifieurs $\mathcal{G}(\mathcal{X}, \mathcal{Y})$ par un espace convexe $\mathcal{H}(\mathcal{X}, \mathbb{R}^K)$ donné par

$$\mathcal{H}(\mathcal{X}, \mathbb{R}^K) = \left\{h = (h^1, \ldots, h^K) : \mathcal{X} \to \mathbb{R}^K\right\}. \tag{1.14}$$

En conséquence, on considère les classifieurs $g_h \in \mathcal{G}(\mathcal{X}, \mathcal{Y})$ définis à partir des fonctions score $h \in \mathcal{H}(\mathcal{X}, \mathbb{R}^K)$ tels que

$$g_h(X) = \underset{i \in \mathcal{Y}}{\arg\max}\ h^i(X). \tag{1.15}$$

Finalement, le nouveau classifieur empirique $g_{\widehat{h}}$ dépend de la fonction score $\widehat{h}$ donnée par

$$\widehat{h} = \underset{h \in \mathcal{H}(\mathcal{X}, \mathbb{R}^K)}{\arg\min}\ \widehat{\mathfrak{R}}_N(h) \tag{1.16}$$

avec pour chaque $h \in \mathcal{H}(\mathcal{X}, \mathbb{R}^K)$,

$$\widehat{\mathfrak{R}}_N(h) = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\left(1 - Z_i h^i(X_j)\right)^2 \tag{1.17}$$

et pour chaque $i \in \mathcal{Y}$, $Z_i = 2\mathbb{1}_{Y=i} - 1 \in \{-1, 1\}$. $\widehat{\mathfrak{R}}_N$ est la version empirique du nouveau risque de classification $\mathfrak{R}$ donné pour tout $h \in \mathcal{H}\left(\mathcal{X}, \mathbb{R}^K\right)$ par

$$\mathfrak{R}(h) = \mathbb{E}\left[\sum_{i=1}^{K}\left(1 - Z_i h^i(X)\right)^2\right]. \tag{1.18}$$

Pour tout $X \in \mathcal{X}$ et pour chaque étiquette $i \in \mathcal{Y}$, la variable aléatoire Z_i vaut $Z_i = 1$ si $g_h(X) = i$ et $Z_i = -1$ si $g_h(X) \neq i$.

Nous présentons dans les chapitres 3 et 4 des procédures de classification multiclasses pour des données de diffusion, en utilisant respectivement le principe *plug-in* et la minimisation du risque empirique de classification.

1.3.5 Modèle de classification étudié dans la thèse

On étudie le modèle de diffusion (1.1) avec $T = 1$. Sous l'hypothèse 1.2.2, le modèle (1.1) admet une unique solution forte (voir par exemple [63]).

On définit pour tout $t \in [0, 1]$, une procédure de classification $g_t : \mathcal{X}_t \longrightarrow \mathcal{Y}$ à l'instant t, avec $\mathcal{X}_t$ l'espace des processus de diffusion X donné par

$$\mathcal{X}_t = (C([0, t], \mathbb{R}), \mathcal{C}_t)$$

où, pour tout $t \in [0, T]$, $C([0, t], \mathbb{R})$ est l'espace des fonctions continues sur l'intervalle $[0, t]$ et à valeurs dans $\mathbb{R}$, et $\mathcal{C}_t$ est sa tribu associée. Ainsi, le classifieur g_t s'actualise lorsque le temps t parcourt l'intervalle $[0, T]$. Dans ce manuscrit, nous nous focalisons sur les classifieurs $g = g_1$ définis à l'horizon temporel $T = 1$, et nous posons $\mathcal{X} = (C([0, 1], \mathbb{R}), \mathcal{C}_1)$.

Le classifieur de Bayes $g^* : \mathcal{X} \longrightarrow \mathcal{Y}$ est toujours donné pour tout $X \in \mathcal{X}$ par

$$g^*(X) = \arg\max_{i \in \mathcal{Y}} \pi_i^*(X) \tag{1.19}$$

où X désigne un processus de diffusion dont la loi de probabilité dépend de l'étiquette $Y \in \mathcal{Y}$.

On fait l'hypothèse supplémentaire suivante.

Hypothèse 1.3.3. *Pour tout $i \in \mathcal{Y}$, nous avons*

$$\mathbb{E}_X \left[\exp\left(\frac{1}{2} \int_0^1 \frac{b_i^{*2}}{\sigma^{*2}}(X_s) ds \right) \right] < +\infty.$$

Cette hypothèse est connue sous le nom de condition de Novikov, et nous servira en particulier à appliquer le théorème de Girsanov pour le changement de probabilité des processus de diffusion. Le résultat ci-dessous, prouvé dans [29], explicite les probabilités conditionnelles $\pi_i^*(X)$ pour chaque classe $i \in \mathcal{Y}$.

Proposition 1.3.4 (Denis, Dion, Martinez (2020))**.** *Supposons que les fonctions de dérive b_i^* et le coefficient de diffusion σ^* satisfont les hypothèses 1.2.2 et 1.3.3. Pour tout $i \in \mathcal{Y}$, on définit la quantité*

$$F_i^*(X) := \int_0^1 \frac{b_i^*}{\sigma^{*2}}(X_s) dX_s - \frac{1}{2} \int_0^1 \frac{b_i^{*2}}{\sigma^{*2}}(X_s) ds.$$

Pour tout $i \in \mathcal{Y}$, la probabilité conditionnelle $\pi^*(i, X)$ est donnée comme suit :

$$\pi_i^*(X) = \phi_i^* \left(\mathbf{F}^*(X) \right),$$

où $\mathbf{F}^* = (F_1^*, \ldots, F_k^*)$, et $\phi_i^* : (x_1, \cdots, x_K) \mapsto \frac{\mathfrak{p}_i^* e^{x_i}}{\sum_{k=1}^K \mathfrak{p}_k^* e^{x_k}}$ sont des fonctions dites softmax.

La proposition 1.3.4 calcule ainsi, pour toute diffusion $X \in \mathcal{X}$, la probabilité $\pi_i^*(X)$ d'appartenance de X à chaque classe $i \in \mathcal{Y}$. Ce résultat est obtenu sous les hypothèses 1.2.2 et 1.3.3, grâce au théorème de Girsanov duquel sont tirées les quantités $F_i^*(X)$. Cependant, ayant supposé que les fonctions de dérive b_i^*, $i \in \mathcal{Y}$, le coefficient de diffusion σ^* et la loi discrète $\mathfrak{p}^*$ de Y sont inconnus, les quantités $F_i^*(X)$ et les fonctions softmax ϕ_i^* sont incalculables, ainsi en est-il également du classifieur de Bayes g^*.

Dans ce manuscrit, nous proposons des procédures de classification empiriques basées sur des observations discrètes de la loi de mélange du processus de diffusion $X = (X_t)_{t \in [0,1]}$, solution du modèle 1.1 et de densité de transition $(t, x) \mapsto p_X(t, x)$ donnée pour tout $(t, x) \in [0, 1] \times \mathbb{R}$ par

$$p_X(t, x) := \sum_{i=1}^K \mathfrak{p}_i^* p_{X,i}(t, x) \tag{1.20}$$

où, pour chaque étiquette $i \in \mathcal{Y}$, $p_{X,i}$ est la densité de transition de X conditionnellement à l'évènement $\{Y = i\}$.

Démonstration. Notons pour chaque $i \in \mathcal{Y}$, $\mathbb{P}_i$ la loi de probabilité conditionnelle à $\{Y = i\}$, et $\mathbb{E}_i$ l'espérance associée. Ainsi, la loi de probabilité $\mathbb{P}$ de X est donnée par

$$\mathbb{P} = \sum_{i \in \mathcal{Y}} \mathfrak{p}_i^* \mathbb{P}_i.$$

On note $W^{\mathbb{P}} = (W_t^{\mathbb{P}})_{t \geq 0}$ le mouvement brownien sous la loi $\mathbb{P}$. Soit $\mathbb{Q}$ une autre loi de probabilité sous laquelle le processus de diffusion X est solution de l'équation $dX_t = \sigma^*(X_t) dW_t^{\mathbb{Q}}$ où, pour tout $t \in [0,1]$, $W_t^{\mathbb{Q}} = W_t^{\mathbb{P}} + \displaystyle\int_0^t \frac{b_Y^*}{\sigma^*}(X_s) ds$.

Le processus $X = (X_t)_{t \in [0,1]}$ étant prévisible car adapté par rapport à sa tribu naturelle $\mathcal{F}^X = (\mathcal{F}_t^X)_{t \in [0,1]}$, alors, d'après le théorème de Girsanov (voir [81]), on a :

i) $\mathbb{Q}$ est équivalente à $\mathbb{P}$.

ii) Pour tout $i \in \mathcal{Y}$, on a

$$\psi_i = \frac{d\mathbb{P}_i|_{\mathcal{F}_1^X}}{d\mathbb{Q}|_{\mathcal{F}_1^X}} = \exp\left(\int_0^1 \frac{b_i^*}{\sigma^*}(X_s) dW_s^{\mathbb{P}_i} + \frac{1}{2} \int_0^1 \frac{b_i^{*2}}{\sigma^{*2}}(X_s) ds \right).$$

iii) $W^{\mathbb{Q}}$ est un mouvement brownien sous $\mathbb{Q}$.

Pour $i \in \mathcal{Y}$, nous avons :

$$\forall\, t \geq 0 \;\; W_t^{\mathbb{P}_i} = W_t^{\mathbb{Q}} - \int_0^t \frac{b_i^*}{\sigma^*}(X_s) ds \;\; et \;\; dW_t^{\mathbb{Q}} = \frac{1}{\sigma^*(X_t)} dX_t.$$

Donc :

$$\int_0^1 \frac{b_i^*}{\sigma^*}(X_s) dW_s^{\mathbb{P}_i} = \int_0^1 \frac{b_i^*}{\sigma^*}(X_s) \left(dW_s^{\mathbb{Q}} - \frac{b_i^*}{\sigma^*}(X_s) \right) ds = \int_0^1 \frac{b_i^*}{\sigma^{*2}}(X_s) dX_s - \int_0^1 \frac{b_i^{*2}}{\sigma^{*2}}(X_s) ds.$$

On en déduit que pour $i \in \mathcal{Y}$:

$$\psi_i = \frac{d\mathbb{P}_i|_{\mathcal{F}_1^X}}{d\mathbb{Q}|_{\mathcal{F}_1^X}} = \exp\left(\int_0^1 \frac{b_i^*}{\sigma^{*2}}(X_s) dX_s - \frac{1}{2} \int_0^1 \frac{b_i^{*2}}{\sigma^{*2}}(X_s) ds \right) = \exp(F_i^*(X)). \tag{1.21}$$

De plus, on a :

$$d\mathbb{P}|_{\mathcal{F}_1^X} = \sum_{i \in \mathcal{Y}} \mathfrak{p}_i^* d\mathbb{P}_i|_{\mathcal{F}_1^X} = \left(\sum_{i \in \mathcal{Y}} \mathfrak{p}_i^* \psi_i \right) d\mathbb{Q}|_{\mathcal{F}_1^X} = \left(\sum_{i \in \mathcal{Y}} \mathfrak{p}_i^* \exp\left(F_i^*(X)\right) \right) d\mathbb{Q}|_{\mathcal{F}_1^X}. \tag{1.22}$$

On déduit alors des résultats 1.21 et 1.22 que :

$$\frac{d\mathbb{P}_i|_{\mathcal{F}_1^X}}{d\mathbb{P}|_{\mathcal{F}_1^X}} = \frac{\exp\left(F_i^*(X)\right)}{\sum_{i \in \mathcal{Y}} \mathfrak{p}_i^* \exp\left(F_i^*(X)\right)}. \tag{1.23}$$

L'objectif est de calculer, pour tout $i \in \mathcal{Y}$ l'espérance conditionnelle $\mathbb{P}(Y = i|\mathcal{F}_1^X) = \mathbb{E}\left(\mathbf{1}_{\{Y=i\}}|\mathcal{F}_1^X\right)$ qui est $\mathcal{F}_1^X-$mesurable et positive. Nous utiliserons pour cela l'orthogonalité de l'espérance conditionnelle. Soit donc Z une variable aléatoire $\mathcal{F}_1^X-$mesurable et bornée. On a :

$$\begin{aligned} \mathbb{E}\left[\mathbb{1}_{\{Y=i\}} Z\right] &= \mathbb{E}\left[\mathbb{E}\left(\mathbb{1}_{\{Y=i\}} Z|Y\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left(\mathbb{1}_{\{Y=i\}} Z|Y=i\right)\right] \\ &= \mathbb{E}\left[\mathbb{1}_{\{Y=i\}} \mathbb{E}_i(Z)\right]. \end{aligned}$$

D'après l'équation (1.23), on obtient

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\{Y=i\}}Z\right] &= \mathbb{E}\left[\mathbb{1}_{\{Y=i\}}\mathbb{E}\left(\frac{d\mathbb{P}_i|_{\mathcal{F}_1^X}}{d\mathbb{P}|_{\mathcal{F}_1^X}}Z\right)\right]\\ &= \mathbb{E}\left[\mathbb{1}_{\{Y=i\}}\right]\mathbb{E}\left[\frac{d\mathbb{P}_i|_{\mathcal{F}_1^X}}{d\mathbb{P}|_{\mathcal{F}_1^X}}Z\right]\\ &= \mathbb{E}\left[\frac{\mathfrak{p}_i^*\exp\left(F_i^*(X)\right)}{\sum_{k\in\mathcal{Y}}\mathfrak{p}_k^*\exp\left(F_k^*(X)\right)}Z\right].\end{aligned}$$

On en déduit finalement que pour $i \in \mathcal{Y}$:

$$\pi_i^*(X) = \mathbb{P}(Y = i|X) = \phi_i^*(\mathbf{F}^*(X))$$

avec $\mathbf{F}^* = (F_1^*, \ldots, F_K^*)$, et $\phi_i^* : (x_1, \ldots, x_K) \mapsto \dfrac{\mathfrak{p}_i^* e^{x_i}}{\sum_{k=1}^K \mathfrak{p}_k^* e^{x_k}}$. □

Nous construisons dans les chapitres 3 et 4 des procédures empiriques de classification imitant le classifieur de Bayes g^*. Cette construction repose soit sur l'estimation non-paramétrique des fonctions b_i^*, $i \in \mathcal{Y}$, σ^{*2} et l'estimation de la loi $\mathfrak{p}^*$, soit sur la minimisation du risque empirique sur des espaces de dimensions finies contenant les éléments inconnus $\mathbf{b}^* = (b_1^*, \ldots, b_K^*)$, σ^{*2} et $\mathfrak{p}^*$. Nous présentons dans la section suivante, des classifieurs pour les données fonctionnelles proposés dans la littérature.

1.3.6 Classification supervisée des données fonctionnelles

La classification supervisée de données fonctionnelles est assez étudiée de manière générale dans la littérature scientifique.

Vue d'ensemble sur la classification de données fonctionnelles. On trouve dans la littérature scientifique des procédures de classification pour divers types de données fonctionnelles. Parmi les plus répandus, nous avons les classifieurs basés sur la notion de profondeur d'une donnée fonctionnelle dans une classe de données. On entend par profondeur pour les données, une fonction qui prend en argument une donnée multivariée ou fonctionnelle, et renvoie une quantité numérique qui détermine le degré de centralité de la donnée par rapport à une plage d'autres données. Cette mesure de profondeur peut être basée sur une mesure de distance. Ainsi, il y a des classifieurs qui prédisent les classes sur la base de la maximisation de la profondeur (voir par exemple [23], [68], [50]). Comme mesures de profondeur, nous avons par exemple la *Mahalanobis depth* (voir par exemple [12], [50]), la *Half-space depth* (voir par exemple [90]), d'autres profondeurs sont basées sur la réduction de dimension des données fonctionnelles (voir par exemple [73]). Nous trouvons ensuite des classifieurs résultant de l'adaptation de l'analyse discriminante linéaire aux données fonctionnelles (voir par exemple [62]), des classifieurs basés sur la méthode des noyaux (voir par exemple [38]), les classifieurs utilisant la méthode des $k-$plus proches voisins (voir par exemple [10]), ou celles basées sur la notion de signature pour des données temporelles (voir par exemple [37]) ou encore les règles de classification construites à l'aide de réseaux de neurones (voir par exemple [40]), la liste est bien-sûr non exhaustive.

Classification de trajectoires de diffusion. L'étude de la classification de trajectoires de diffusion n'est pas encore assez répandue dans la littérature, les données modélisées par des équations différentielles stochastiques étant de plus en plus utilisées dans des domaines tels que la finance, ou encore la physique. Le but est de proposer une procédure de classification qui s'appuie sur les propriétés des processus de diffusion pour classer avec plus d'efficacité des données issues de ce type de processus stochastique, comparé aux procédures de classification adaptées à toutes données fonctionnelles. Nous trouvons dans la littérature les références ci-dessous.

1. **Gadat, Gerchinovitz, Marteau (2020) [44].**

Cet article propose une procédure de classification binaire de type *plug-in* pour les trajectoires de diffusion solutions du modèle de bruit blanc

$$dX_t = Y f(t)dt + (1-Y)g(t)dt + dW_t, \quad t \in [0,1] \tag{1.24}$$

où $(W_t)_{t\geq 0}$ est un mouvement brownien standard, Y est une variable aléatoire suivant une loi de Bernoulli de paramètre $1/2$ indépendante de $(W_t)_{t\geq 0}$, et les fonctions $f, g \in \mathbb{L}^2([0,1])$. Comme le coefficient de diffusion $\sigma^* = 1$ et les fonctions de dérive f et g dépendent du temps, il en résulte que le processus X solution du modèle (1.24) est un processus gaussien qui satisfait pour tout $t \in (0,1]$,

$$X_t|Y=0 \sim \mathcal{N}(\nu_t, 1), \;\; X_t|Y=1 \sim \mathcal{N}(\mu_t, 1),$$

où

$$\nu_t = \int_0^t g(s)ds, \quad \mu_t = \int_0^t f(s)ds.$$

Construction du classifieur empirique. Rappelons que le classifieur de Bayes noté Φ^* est défini par

$$\Phi^* \in \arg\min_{\Phi} \mathcal{R}(\Phi),$$

avec $\mathcal{R}(\Phi) = \mathbb{P}(\Phi(X) \neq Y)$ le risque de mauvaise classification. Le classifieur Φ^* est caractérisé pour tout processus gaussien X solution de (1.24) par

$$\Phi^*(X) := \mathbb{1}_{\eta^*(X)\geq 1/2},$$

avec $\eta^*(X) = \mathbb{P}(Y=1|X)$ la fonction de régression associée au modèle étudié. Il est ensuite établi que

$$\Phi^*(X) = \mathbb{1}_{\eta^*(X)\geq 1/2} = \mathbb{1}_{\left\{\int_0^1 (f(s)-g(s))dX_s \geq \frac{1}{2}\|f\|^2 - \frac{1}{2}\|g\|^2\right\}}.$$

Les auteurs proposent une procédure de classification de type *plug-in* basée sur des estimateurs par projection des fonctions de dérive f et g sur un sous-espace de $\mathbb{L}^2([0,1])$ de dimension finie, à partir d'un échantillon de données $\{(X^k, Y_k), k = 1, \ldots, N\}$ constitué de N copies indépendantes du couple aléatoire (X, Y).

En effet, soient

$$f_d := \Pi_d(f) = \sum_{j=1}^{d} \theta_j \phi_j \;\text{ et }\; g_d := \Pi_d(g) = \sum_{j=1}^{d} \mu_j \phi_j$$

où Π_d est un projecteur orthogonal sur un sous-espace vectoriel de $\mathbb{L}^2([0,1])$ de dimension d, généré par une base orthonormal $(\phi_j)_{j=1,\ldots,d}$. Ils construisent ensuite, à partir de l'echantillon de trajectoires de diffusions $\{(X^k, Y_k), k = 1, \ldots, N\}$, l'échantilon de données univariées

$$\mathcal{E} = \left\{\left\langle \phi_1, X^1 \right\rangle, \ldots, \left\langle \phi_1, X^N \right\rangle, \left\langle \phi_2, X^1 \right\rangle, \ldots, \left\langle \phi_2, X^N \right\rangle, \ldots, \left\langle \phi_d, X^1 \right\rangle, \ldots, \left\langle \phi_d, X^N \right\rangle\right\}$$

avec pour tous $j = 1, \ldots d$ et $k = 1, \ldots, N$,

$$\left\langle \phi_j, X^k \right\rangle = \int_0^1 \phi_j(t) dX_t^k.$$

L'échantillon initial $\mathcal{E}$ est ensuite divisé en deux sous-échantillons

$$\mathcal{E}_0 = \left\{\left\langle \phi_j, X^{k,0} \right\rangle, \;\; j = 1, \ldots, d, \;\; k = 1, \ldots, N_0\right\},$$

avec $N_0 := \sum_{k=1}^{N} \mathbb{1}_{Y_k=0}$, et

$$\mathcal{E}_1 = \left\{\left\langle \phi_j, X^{k,1} \right\rangle, \;\; j = 1, \ldots, d, \;\; k = 1, \ldots, N_1\right\},$$

avec $N_1 := \sum_{k=1}^{N} \mathbb{1}_{Y_k=1}$ correspondant respectivement aux classes $Y = 0$ et $Y = 1$. Enfin, les fonctions f et g sont estimées respectivement par $\widehat{f}_d = \sum_{j=1}^{d} \widehat{\theta}_j \phi_j$ et $\widehat{g}_d = \sum_{j=1}^{d} \widehat{\mu}_j \phi_j$ avec pour tout $j = 1, \ldots, d$,

$$\widehat{\theta}_j = \mathbb{1}_{N_1>0} \frac{1}{N_1} \sum_{k=1}^{N_1} \left\langle \phi_j, X^{k,1} \right\rangle, \quad \widehat{\mu_j} = \mathbb{1}_{N_0>0} \frac{1}{N_0} \sum_{k=1}^{N_0} \left\langle \phi_j, X^{k,1} \right\rangle.$$

Vitesse de convergence du classifieur empirique. L'étude de la vitesse de convergence du classifieur empirique est réalisée en tirant profit du caractère gaussien du processus X et du fait que les fonctions de dérive f et g sont à support compact. Une hypothèse de marge est faite sur la fonction de régression η^*, impliquant une condition de séparation des classes caractérisée par

$$\Delta = \|f - g\| = \sqrt{\int_0^1 (f-g)^2(t)dt}.$$

Ils établissement ainsi, en supposant que f et g appartiennent à la boule de Sobolev $\mathcal{H}_s(R)$ de paramètre de régularité s, avec $R > 0$, et en posant $d = d_N = \lfloor (R^2 N)^{1/(2s+1)} \rfloor$, une vitesse de convergence de l'ordre de $N^{-s/(2s+1)} \log(N)$ si $\Delta < R^{1/(2s+1)} N^{-s/(2s+1)} \log(N)$, et une vitesse de l'ordre de $\frac{1}{\Delta} N^{-2s/(2s+1)} \log^2(N)$ si $\Delta \geq R^{1/(2s+1)} N^{-s/(2s+1)} \log(N)$. Il est remarqué qu'une distance de séparation large entraine une vitesse de convergence beaucoup plus rapide.

Dans le chapitre 4, nous étendons le modèle étudié au modèle de diffusion dont les fonctions de dérive dépendent de l'espace et sont définies sur la droite réelle $\mathbb{R}$. Nous établissons une vitesse de convergence plus rapide d'un classifieur de type ERM construit en classification binaire, en imposant une hypothèse de marge sur la fonction de régression associée au modèle.

2. **Cadre (2013) [11].**

L'auteur propose une procédure de classification binaire de trajectoires de diffusions issues d'un mélange de diffusions solutions du modèle ci-dessous

$$(M): \begin{cases} dX_t = b_0(t, X_t)dt + \sigma_0(t, X_t)dW_t \ \text{ lorsque } \ Y = 0, \\ \\ dX_t = (b_0(t, X_t) + (f_0\sigma_0)(t, X_t))dt + \sigma_0(t, X_t)dW_t \ \text{ lorsque } \ Y = 1 \end{cases}, \quad (t \in [0,1])$$

avec $(W_t)_{t\in[0,1]}$ un mouvement brownien standard indépendant de de l'étiquette $Y \in \{0, 1\}$, X_0 est indépendante de Y et $(B_t)_{t\in[0,1]}$, et $f_0, b_0, \sigma_0 : [0, 1] \times \mathbb{R} \longrightarrow \mathbb{R}$ sont des fonctions boréliennes telles que le modèle (M) admet une solution forte étant donnée X_0. La variable aléatoire Y est de loi discrète (p_0, p_1) inconnue, avec $p_0 = \mathbb{P}(Y = 0) \in (0, 1)$.

Le classifieur de Bayes g^* est défini pour tout X par $g^*(X) = \mathbb{1}_{\eta^*(X)>1/2}$ avec η^* la fonction de régression associée au modèle et donnée pour tout $x \in \mathbb{R}$ par $\eta^*(x) = \mathbb{P}(Y = 1|X = x)$. Il explicite la formule de la fonction de régression η^* et obtient

$$\eta^*(X) = \frac{(1-p_0)\exp(F(h_0, X))}{p_0 + (1-p_0)\exp(F(h_0, X))}$$

avec,

$$h_0 = \left(\frac{f_0}{\sigma_0}, \frac{f_0 b_0}{\sigma_0} + \frac{f_0^2}{2} \right)$$

et,

$$F(h_0, X) = \int_0^1 \frac{f_0(t, X_t)}{\sigma_0(t, X_t)} dX_t - \int_0^1 \left(\frac{f_0 b_0}{\sigma_0} + \frac{f_0^2}{2} \right)(t, X_t)dt.$$

Il en ressort que pour toute diffusion X solution du modèle (M),

$$g^*(X) = \mathbb{1}_{\eta^*(X)>1/2} = \mathbb{1}_{\left\{F(h_0,X)>\log\left(\frac{p_0}{1-p_0}\right)\right\}}.$$

Construction du classifieur empirique de type ERM. La performance d'un classifieur g est mesurée par son risque de classification $\mathcal{R}(g) = \mathbb{P}(g(X) \neq Y)$, et, comme on le sait déjà, le classifieur de Bayes minimise la fonction $\mathcal{R}$ à valeurs dans $[0,1]$. Comme la loi des observations est supposée inconnue, il considère, à partir d'un échantillon d'apprentissage $(X_j, Y_j)_{j=1,\dots,N}$ le risque empirique de classification $\widehat{\mathcal{R}}$ donné pour tout classifieur g par

$$\widehat{\mathcal{R}}(g) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{g(X_j)\neq Y_j}.$$

Il pose $g^* = g(h_0, p_0)$, et il considère les classifieurs s'écrivant sous la forme $g(h,p)$ avec $h : [0,1]\times\mathbb{R} \longrightarrow \mathbb{R}$ et $p \in [0,1]^2$ une loi de probabilité discrète. Soit $\mathcal{F}$ l'ensemble des fonctions $h : [0,1]\times\mathbb{R} \longrightarrow \mathbb{R}$ et pour $\varepsilon > 0$, $\mathcal{F}_\varepsilon$ son $\varepsilon-$net. on considère ensuite l'ensemble discret $I_\varepsilon = \{[k\varepsilon],\ k = 1,\dots,[1/\varepsilon]\}$. Ainsi, le classifieur empirique est donné par

$$\widehat{g}_\varepsilon(X) = g(\widehat{h}_\varepsilon, \widehat{p}_\varepsilon)(X) = \mathbb{1}_{\left\{F(\widehat{h}_\varepsilon, X) > \log\left(\frac{\widehat{p}_\varepsilon}{1-\widehat{p}_\varepsilon}\right)\right\}}$$

avec

$$(\widehat{h}_\varepsilon, \widehat{p}_\varepsilon) = \underset{(h,p)\mathcal{F}_\varepsilon\times I_\varepsilon^2}{\arg\min}\ \widehat{\mathcal{R}}(g(h,p)).$$

Vitesse de convergence du classifieur empirique. Le cardinal de l'$\varepsilon-$net $\mathcal{F}_\varepsilon$ est noté $\mathfrak{N}(\varepsilon)$. Ainsi, sous des hypothèses de régularité sur les fonctions f_0, b_0 et σ_0, et supposant qu'il existe $u > 0$ et $C \geq 1$ tel que pour tout $\varepsilon > 0$, $\log(\mathfrak{N}(\varepsilon)) \leq C\varepsilon^{-u}$, il obtient le résultat suivant

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}_\varepsilon)\right] - \mathcal{R}(g^*) \leq 2\left(\frac{K^u}{n}\right)^{1/(2+u)}\left(\frac{2}{u} + C\right) + \sqrt{\frac{\log(\Delta)}{n}} + \frac{1}{\Delta\sqrt{n\Delta}}$$

avec,

$$\Delta = \sum_{\varepsilon\in\mathcal{E}} \frac{\mathfrak{N}(\varepsilon)}{\varepsilon}\exp(-2\lambda_\varepsilon^2) < \infty,$$

où $\mathcal{E} = \{1/\ell,\ \ell \geq 1\}$, et $\lambda_\varepsilon > 0$ un terme de pénalité choisi tel que

$$\lambda_\varepsilon^2 \geq 2\log\left(\frac{1}{\varepsilon}\right) + \log(\mathfrak{N}(\varepsilon)).$$

Lorsqu'il considère par exemple le cas des diffusions homogènes en temps, il obtient, sous des hypothèses de régularité caractérisées par un paramètre $k \geq 1$ sur les fonctions f_0, b_0 et σ_0, que $\log(\mathfrak{N}(\varepsilon)) = \mathrm{O}(\varepsilon^{-1/k})$, et

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}_\varepsilon)\right] - \mathcal{R}(g^*) = \mathrm{O}\left(N^{-k/(2k+1)}\right).$$

3. **Denis, Dion, Martinez (2020)** [29].

   Cet article propose d'une part, une procédure paramétrique de classification multiclasses de type plug-in pour les trajectoires de diffusion. D'autre part, ils proposent une procédure non-paramétrique de classification multiclasses de type ERM pour les données de diffusion. Ces données sont générées par un mélange de processus de diffusion avec des fonctions de dérive inconnues et dépendant des classes, et un coefficient de diffusion connu et commun à toutes les classes. Ils établissent une vitesse de convergence du classifieur de type ERM de l'ordre de $N^{-\beta/(2\beta+1)}$ sur un espace de Hölder de régularité $\beta \geq 1$. Toutefois, le classifieur ainsi construit n'est pas calculable en pratique, étant solution d'un problème d'optimisation non-convexe.

La procédure de classification de type *plug-in* proposée dans ce manuscrit étend le cadre d'étude dans [29] à un cadre non-paramétrique. Le classifieur de type ERM est quant à lui implémentable contrairement à celui proposé dans [11], et il est étendu au cadre multiclasses. On s'inspire ensuite

de l'hypothèse de marge sur la fonction de régression en classification binaire décrite dans [44] pour proposer un classifieur de type ERM avec une vitesse plus rapide.

Nous présentons dans la section suivante les principales contributions de la thèse sur le problème d'estimation non-paramétrique du coefficient de diffusion, et sur la construction des procédures non-paramétriques de classification pour les trajectoires de diffusions.

1.4 Contributions de la thèse

Nous proposons dans ce manuscrit des procédures de classification pour des trajectoires de diffusion dans un cadre plus général où les coefficients de dérive et de diffusion sont inconnus ainsi que la loi discrète de l'étiquette $Y \in \mathcal{Y}$. Ce travail est réparti en trois principaux chapitres. Le chapitre 2 est consacré à l'estimation non-paramétrique du coefficient de diffusion d'une équation différentielle stochastique homogène en temps. Dans le chapitre 3, nous proposons une procédure non-paramétrique de classification multiclasses de type *plug-in* pour les trajectoires de diffusions. Nous proposons ensuite au chapitre 4 une procédure de classification multiclasses basée sur la minimisation du risque empirique de classification.

1.4.1 Estimation non-paramétrique du coefficient de diffusion

Modèle et notations. On considère $X = (X_t)_{t\in[0,1]}$ un processus stochastique solution du modèle (1.1). On suppose que le processus est observé en temps discret à haute fréquence sur l'intervalle de temps $[0, 1]$.

Nous proposons un estimateur non-paramétrique de σ^{*2} en distinguant d'une part, le cas où une seule trajectoire de diffusion $\bar{X} = (X_{k\Delta_n})$ est observée, avec $\Delta_n = 1/n$, et d'autre part, le cas où on suppose disposer d'un échantillon $(\bar{X}^1, \ldots, \bar{X}^N)$ constitué de N copies indépendantes de $\bar{X}$ avec $N, n \to \infty$. Nous proposons des estimateurs non-paramétriques de σ^{*2} respectivement sur un intervalle compact, et sur la droite réelle $\mathbb{R}$, et enfin, nous établissons des vitesses de convergence dans chacun des cas.

Construction des estimateurs. Nous approximons l'espace de fonctions lipschitziennes par un sous-espace vectoriel $\mathcal{S}_m$ de dimension finie $m \geq 1$ généré soit par la base de fonctions **B**-splines

$$\{B_\ell, \ell = 0, \ldots, m-1\}$$

construite sur un intervalle $I = [-A, A]$ de $\mathbb{R}$ avec $A > 0$ (voir par exemple [54]), soit par une base orthonormée comme par exemple la base de Fourier, ou encore la base de Hermite.

On considère le sous-espace contraint $\mathcal{S}_{m,L}$ défini par

$$\mathcal{S}_{m,L} := \left\{ h = \sum_{\ell=0}^{m-1} a_\ell f_\ell \in \mathcal{S}_m : \sum_{\ell=0}^{m-1} a_\ell^2 \leq mL \right\}$$

où $L \in \mathbb{R}_{+*}$. Soit $\gamma_{n,N}$ la fonction de contraste des moindres carrés définie pour tout $h \in \mathcal{S}_{m,L}$ par

$$\gamma_{n,N}(h) := \frac{1}{Nn} \sum_{j=1}^{N} \sum_{k=0}^{n-1} \left(U_{k\Delta}^j - h(X_{k\Delta_n}^j) \right)^2$$

où pour tout $j \in [\![1, N]\!]$ et pour tout $k \in [\![0, n-1]\!]$, $U_{k\Delta_n}^j := (X_{(k+1)\Delta_n}^j - X_{k\Delta_n}^j)^2/\Delta_n$.

Ainsi, pour chaque dimension $m \geq 1$, l'estimateur non-paramétrique $\widehat{\sigma}_m^2$ de σ^{*2} est donné par

$$\widehat{\sigma}_m^2 \in \underset{h\in\mathcal{S}_{m,L}}{\arg\min} \gamma_{n,N}(h). \tag{1.25}$$

L'équation (1.25) équivaut à $\widehat{\sigma}_m^2 = \sum_{\ell=0}^{m-1} \widehat{a}_\ell B_\ell$ avec

$$\widehat{\mathbf{a}} = (\widehat{a}_{-M}, \ldots, \widehat{a}_{K-1})' := \underset{\|a\|_2^2 \leq mL}{\arg\min} \|\mathbf{U} - \mathbf{F}_m \mathbf{a}\|_2^2 \tag{1.26}$$

où $\mathbf{U} = \left(U^1_{0\Delta_n}, \ldots, U^1_{(n-1)\Delta_n}, \ldots, U^N_{0\Delta_n}, \ldots, U^N_{(n-1)\Delta_n}\right)'$ et la matrice $\mathbf{F}_m$ est donnée par

$$\mathbf{F}_m := \left((B_\ell(X^j_0), \ldots, B_\ell(X^j_{(n-1)\Delta_n}))'\right)_{\substack{0\leq \ell \leq m-1 \\ 1\leq j\leq N}} \in \mathbb{R}^{Nn\times m}.$$

L'estimateur $\widehat{\mathbf{a}}$ (ou $\widehat{\sigma}^2_m$) est dit de type ridge grâce à la contrainte ℓ^2 donnée dans l'équation (1.26).

Vitesses de convergence des estimateurs non-adaptatifs. Nous proposons des estimateurs ridge de $\sigma^{*2}_{|I}$, où l'intervalle d'estimation est soit compact (par exemple $I = [-1, 1]$), soit la droite réelle ($I = \mathbb{R}$). On précise que la dimension m du sous-espace d'approximation $\mathcal{S}_{m,L}$ dépend de la taille N de l'échantillon et la discrétisation de l'intervalle de temps $[0, 1]$ (ou longueur n des trajectoires) lorsque $N \to \infty$, ou uniquement de la discrétisation du temps lorsque qu'une seule trajectoire est observée.

Pour tout $h \in \mathcal{S}_{m,L}$, on définit les normes et pseudo-normes suivantes

$$\|h\|^2_{n,N} = \frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1} h^2(X^j_{k\Delta_n}), \quad \|h\|^2_n = \mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1} h^2(X_{k\Delta_n})\right],$$
$$\|h\|^2_X := \int_0^1 h^2(X_s)ds, \quad \|h\|^2 := \int_I h^2(x)dx.$$

On étudie la performance de l'estimateur $\widehat{\sigma}^2_m$ en regardant principalement son risque d'estimation $\mathbb{E}\left[\left\|\widehat{\sigma}^2_m - \sigma^{*2}_{|I}\right\|^2_{n,N}\right]$. On pourra aussi, dans certains cas, considérer le risque d'estimation $\mathbb{E}\left\|\widehat{\sigma}^2_m - \sigma^{*2}_{|I}\right\|^2_n$ défini avec la norme empirique $\|.\|_n$.

Estimation non-paramétrique à partir d'une seule trajectoire. L'obtention des résultats théoriques pour l'estimateur de $\sigma^{*2}_{|I}$, lorsque l'intervalle d'estimation I est un compact, est basée sur le temps $x \mapsto L^x$ donné par l'équation (1.7). Nous obtenons le premier résultat suivant.

Theorem 1.4.1. *Soit $\gamma \in [2, +\infty)$. Supposons que I est compact. Sous des hypothèses sur b^* et σ^*, il existe une constante $C > 0$ dépendant de σ^*_1 telle que*

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_m - \sigma^{*2}_I\right\|^2_{n,1}\right] \leq 3\inf_{h\in\mathcal{S}_{m,L}}\left\|h - \sigma^{*2}_I\right\|^2_n + C\left(\frac{m}{n} + \frac{m^{2\gamma+1}L}{n^{\gamma/2}} + \Delta^2_n\right). \tag{1.27}$$

Les deux premiers termes du membre de droite de l'équation (1.27) sont respectivement le terme de biais et le terme de variance. Le paramètre $\gamma \geq 2$ qui résulte de l'inégalité de Hölder, dépend de la régularité de l'espace d'appartenance de la fonction σ^{*2} et est choisi de sorte que le terme $m^{2\gamma+1}L/n^{\gamma/2}$ soit négligeable devant l'erreur d'estimation m/n. Le troisième terme resulte de la relation d'équivalence entre les pseudo-normes $\|.\|_{n,1}$ et $\|.\|_X$. Le dernier terme Δ^2_n (où $\Delta_n = 1/n$) représente le coût de la discrétisation du temps. Soit $\Sigma_I(\beta, R)$ l'espace de fonctions Hölder donné par

$$\Sigma_I(\beta, R) := \left\{h \in \mathcal{C}^{\lfloor\beta\rfloor+1}(I), \ \left|h^{(\ell)}(x) - h^{(\ell)}(y)\right| \leq R|x-y|^{\beta-l}, \ \ x, y \in I\right\},$$

où $\ell = \lfloor\beta\rfloor$, $\beta \geq 1$ est le paramètre de régularité et $R \in \mathbb{R}_{+*}$. Nous déduisons le résultat suivant.

Corollaire 1.4.2. *Supposons que $\sigma^2 \in \Sigma_I(\beta, R)$ avec $\beta > 3/2$, et l'espace d'approximation est généré par les fonctions **B**-splines. Sous des hypothèses sur L et m,*

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m_{\text{opt}}} - \sigma^{*2}_I\right\|^2\right] = \mathrm{O}\left(n^{-2\beta/(2\beta+1)}\right).$$

La vitesse obtenue est du même ordre que celle établie dans [58] sur une boule de Besov. Le paramètre γ donné dans le théorème 1.4.1 vaut $\gamma = 4(2\beta+1)/(2\beta-3)$, ce qui implique que l'erreur d'estimation est de l'ordre de m/n.

Nous étendons ensuite l'étude au cas où l'intervalle d'estimation est la droite réelle $I = \mathbb{R}$, en posant cette fois $\sigma^{*2}_{|I} = \sigma^{*2}$. Nous tronquons l'estimateur $\widehat{\sigma}^2_m$ donné par les équations (1.25) et (1.26) et nous obtenons l'estimateur $\widehat{\sigma}^2_{m,L}$ défini pour tout $x \in \mathbb{R}$ par

$$\widehat{\sigma}^2_{m,L}(x) = \widehat{\sigma}^2_m(x)\mathbb{1}_{\widehat{\sigma}^2_m(x)\leq\sqrt{L}} + \sqrt{L}\mathbb{1}_{\widehat{\sigma}^2_m(x)>\sqrt{L}}. \tag{1.28}$$

Cette troncature sert particulièrement à optimiser l'ordre de l'erreur d'estimation de σ^{*2} sur $\mathbb{R}$. Nous obtenons le résultat de consistance ci-dessous.

Theorem 1.4.3. *Sous des hypothèses sur b^*, σ^*, m et L, il existe une constante $C > 0$ dépendant de β et σ_1 telle que*

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L} - \sigma^{*2}\right\|^2_{n,1}\right] \leq C\log^{2\beta}(n)n^{-2\beta/(4\beta+1)}.$$

Cette vitesse est, comme attendu, plus lente que celle établie pour l'estimation de $\sigma^{*2}\mathbb{1}_I$ avec I un compact de $\mathbb{R}$. Toutefois, l'optimalité de cette vitesse n'est pas établie dans ce manuscrit.

Estimation non-paramétrique à partir d'observations répétées. Nous étendons l'étude au cas d'observations répétées, disposant de l'échantillon $(\bar{X}^1, \ldots, \bar{X}^N)$ avec $N \to \infty$. Dans ce contexte, il n'est plus nécessaire d'avoir recours au temps local du processus de diffusion pour l'étude théorique de la performance de l'estimateur $\widehat{\sigma}^2_m$. En effet, nous tirons profit de l'indépendance des copies $\bar{X}^j$, $j = 1, \ldots, N$ du processus X pour établir la convergence de l'estimateur. D'autre part, nous utilisons la fonction f_n, définie à partir de la densité de transition p_X du processus X par

$$f_n(x) = \frac{1}{n}\sum_{k=1}^{n-1} p_X(k\Delta, x), \tag{1.29}$$

pour obtenir l'équivalence entre la norme empirique $\|.\|_n$ et la norme L^2 $\|.\|$ établie dans [30]. En effet, si l'intervalle d'estimation I est compact, alors il existe une constante $\tau_0 > 0$ telle que

$$\forall\, x \in I, \quad f_n(x) \geq \tau_0,$$

et on déduit de cette minoration que pour toute fonction continue h,

$$\|h\|^2 \leq \frac{1}{\tau_0}\|h\|^2_n.$$

Nous établissons le résultat de consistance ci-dessous.

Theorem 1.4.4. *Supposons que $\sigma^{*2} \in \Sigma_I(\beta, R)$ avec $\beta > 3/2$ et que l'espace d'approximation est généré par la base de fonctions splines. De plus, supposons que l'intervalle I est compact. Sous des hypothèses sur N, n, m, et L, nous obtenons*

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m_{\text{opt}}} - \sigma^{*2}_I\right\|^2_{n,N}\right] = \mathrm{O}\left((Nn)^{-2\beta/(2\beta+1)}\right).$$

Nous obtenons ainsi une vitesse plus rapide par rapport au cas $N = 1$ en tirant profit des observations répétées du processus de diffusion X.

Nous considérons ensuite le cas où l'intervalle d'estimation $I = \mathbb{R}$, et dans ce cas, la fonction de densité f_n donnée par l'équation (1.29) n'est plus minorée inférieurement. Il devient dès lors plus fastidieux d'établir un résultat de consistance de l'estimateur non-paramétrique de σ^{*2}. Nous considérons une fois de plus, l'estimateur tronqué $\widehat{\sigma}^2_{m,L}$ donné par (1.28), dans le cas où $I = \mathbb{R}$. Nous obtenons alors le résultat ci-dessous.

Theorem 1.4.5. *Supposons que $\sigma^{*2} \in \Sigma_I(\beta, R)$ avec $\beta \geq 1$, $I = \mathbb{R}$, et que l'espace d'approximation est généré par la base de fonctions **B**-splines. Sous des hypothèses sur b^*, σ^*, m et L, on obtient*

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L} - \sigma^{*2}\right\|^2_{n,N}\right] \leq \mathrm{O}\left(\log^{2\beta}(N)(Nn)^{-2\beta/(4\beta+1)} + \frac{1}{n^2}\right)$$

On peut remarquer, d'après les résultats des théorèmes 1.4.4 et 1.4.5, que l'estimateur de σ^{*2} sur l'intervalle $I = \mathbb{R}$ est moins performant comparé à l'estimateur de σ^{*2} sur un intervalle compact.

Estimateur non-adaptatif du coefficient de diffusion sur un intervalle compact dépendant de N. Nous considérons dans cette partie, l'intermédaire entre le cas I compact, et le cas $I = \mathbb{R}$. En effet, on propose un estimateur ridge noté $\widehat{\sigma}^2_{A_N,m}$ de $\sigma^{*2}_{A_N} = \sigma^{*2}_{|I}$ par projection sur la base des splines, sur l'intervalle compact $I = [-A_N, A_N]$ où $A_N > 0$ tend vers l'infini lorsque N tends vers l'infini. Ainsi, grâce au profit tiré des propriétés des fonctions **B**-splines (voir [54]) et à un choix approprié de A_N en fonction de N, nous obtenons le résultat suivant.

Théorème 1.4.6. *Supposons que $N \propto n$ et considérons l'estimateur ridge $\widehat{\sigma}^2_{A_N,m}$ de $\sigma^{*2}_{A_N}$ basé les fonctions **B**-splines. Supposons également que $\sigma^{*2} \in \Sigma_I(\beta, R)$ avec $I = [-A_N, A_N]$. Sous des hypothèses sur b^*, σ^*, L, m et A_N, on obtient*

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{A_N,m} - \sigma^{*2}_{A_N}\right\|^2_{n,N}\right] \leq C \log^{\beta}(N)(Nn)^{-2\beta/(2\beta+1)}$$

où $C > 0$ est une constante dependant de $\beta \geq 1$.

On obtient ainsi une vitesse du même ordre (à un facteur logarithmique près) que la vitesse établie dans le théorème 1.4.4 dans le cas où I est compact et indépendant de N. Nous verrons que l'obtention de ce résultat nécessite toutefois que l'intervalle d'estimation $[-A_N, A_N]$ converge lentement vers $\mathbb{R}$.

Estimateur adaptatif du coefficient de diffusion. Les estimateurs proposés dans la section précédente dépendent de la dimension du sous-espace d'approximation qu'il faut choisir en pratique. Ainsi, on considère l'estimateur adaptatif $\widehat{\sigma}^2_{\widehat{m},L}$ où la dimension $\widehat{m}$ est sélectionnée telle que

$$\widehat{m} := \underset{m\in\mathcal{M}}{\arg\min}\left\{\gamma_{n,N}\left(\widehat{\sigma}^2_m\right) + \mathrm{pen}(m)\right\} \tag{1.30}$$

où $\mathrm{pen} : m \mapsto \kappa m \log(N)/(Nn)$, avec κ à calibrer, est la fonction de pénalité et $\mathcal{M}$ est un ensemble de valeurs possibles de la dimension choisi tel que pour tous $m, m' \in \mathcal{M}$, $m < m'$ implique que $\mathcal{S}_m \subset \mathcal{S}_{m'}$. Cette propriété est nécessaire pour l'établissement du résultat ci-dessous.

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{\widehat{K},L} - \sigma^{*2}_{|I}\right\|^2_{n,N}\right] \leq \mathfrak{c} \inf_{m\in\mathcal{M}}\left\{\inf_{h\in\mathcal{S}_{m,L}} \|h - \sigma^{*2}_{|I}\|^2_n + \mathrm{pen}(m)\right\} + \frac{C}{Nn}$$

où $\mathfrak{c}$ est une constante numérique dont la valeur dépend principalement de l'intervalle d'estimation I. Notons également que le choix de l'ensemble $\mathcal{M}$ dépend du fait que $N = 1$ ou $N \to \infty$.

1.4.2 Classifieur de type *plug-in* pour les trajectoires de diffusion

On considère le modèle de diffusion (1.1) pour la classification multiclasses. On suppose que le processus $X = (X_t)_{t\in[0,1]}$, solution du modèle (1.1) est observé en temps discret à haute fréquence sur l'intervalle de temps $[0, 1]$. La loi des observations étant supposée inconnue, nous partons d'un échantillon d'apprentissage $\mathcal{D}_N = \left((\bar{X}^j, Y_j),\ j = 1, \ldots, N\right)$ de taille N avec pour chaque $j \in [\![1, N]\!]$, $\bar{X}^j = (X^j_{k\Delta_n})_{0\leq k\leq n}$, et $\Delta_n = 1/n$ est le pas de discrétisation. On suppose que la taille de l'échantillon N tend vers l'infini et le pas de discrétisation Δ_n tend vers zéro.

On note $\mathcal{X} = \mathcal{C}([0, 1], \mathbb{R})$ l'ensemble des trajectoires et $\left(\mathcal{F}^X_t\right)_{t\in[0,1]}$ la filtration naturelle du processus $X = (X_t)_{t\in[0,1]}$. Nous proposons dans cette section une procédure de classification empirique consistante qui va imiter le classifieur de Bayes présenté dans la section 1.3.5.

Classifieur empirique de type *plug-in*. Soient $\widehat{\mathbf{b}} = \left(\widehat{b}_1, \ldots, \widehat{b}_K\right)$, $\widehat{\sigma}^2$ et $\widehat{\mathfrak{p}} = \left(\widehat{\mathfrak{p}}_1, \ldots, \widehat{\mathfrak{p}}_K\right)$ des estimateurs respectifs de $\mathbf{b}^* = (b^*_1, \ldots, b^*_K)$, σ^{*2} et $\mathfrak{p}^*$ construits à partir de $\mathcal{D}_N$. Les estimateurs $\widehat{\mathfrak{p}}_i,\ i \in \mathcal{Y}$ sont donnés par

$$\widehat{\mathfrak{p}}_i = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{Y_j=i},\ \ i = 1, \ldots, K.$$

Considérons le classifieur empirique de type *plug-in* qui imite le classifieur de Bayes $g^* = g_{\mathbf{b}^*,\sigma^{*2},\mathfrak{p}}$, noté $\widehat{g}$, tel que

$$\widehat{g} = g_{\widehat{\mathbf{b}},\widehat{\sigma}^2,\widehat{\mathfrak{p}}}.$$

On définit pour toute fonction $h \in \mathbb{L}^2(\mathbb{R})$, une norme empirique dépendant d'une observation discrète $\bar{X} = (X_{k\Delta_n})_{k\in[\![0,n]\!]}$ de X :

$$\|h\|_n^2 := \mathbb{E}_X\left[\frac{1}{n}\sum_{k=0}^{n-1} h^2(X_{k\Delta})\right],$$

avec $\mathbb{E}_X$ l'espérance correspondant à la loi du processus de diffusion X. Le premier résultat ci-dessous donne une majoration de l'excès de risque en fonction des risques d'estimation des coefficients de dérive b_i^*, du coefficient de diffusion σ^{*2} et de la loi discrète $\mathfrak{p}^*$ de l'étiquette Y.

Théorème 1.4.7 ([28]). *Supposons que N et n sont fixés et suffisamment grands, et qu'il existe des constantes $b_{\max}, \sigma_0^2 > 0$ telles que pour tout $x \in \mathbb{R}$*

$$\max_{i\in\mathcal{Y}} |\widehat{b}_i(x)| \le b_{\max} \quad \text{et} \quad \widehat{\sigma}^2(x) \ge \sigma_0^2.$$

Sous certaines hypothèses sur les fonctions b_i^* et σ^*, le classifieur empirique $\widehat{g}$ satisfait

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \le C\left(\sqrt{\Delta_n} + \frac{1}{\mathfrak{p}_0^*\sqrt{N}} + \mathbb{E}\left[b_{\max}\sigma_0^{-2}\sum_{i=1}^{K}\|\widehat{b}_i - b_i^*\|_n\right] + \mathbb{E}\left[\sigma_0^{-2}\|\widehat{\sigma}^2 - \sigma^{*2}\|_n\right]\right),$$

où $C > 0$ est une constante qui dépend de $\mathbf{b}^*$, σ^*, et K.

On déduit du résultat du théorème 1.4.7 ci-dessus que la consistance du classifieur empirique $\widehat{g}$ est étroitement liée à la qualité d'estimation des fonctions de dérive $b_i^*,\ i = 1,\dots,K$ et du carré du coefficient de diffusion σ^{*2}.

Consistance du classifieur empirique $\widehat{g}$. Les fonctions inconnues $b_i^*,\ i \in \mathcal{Y}$ et σ^{*2} sont estimées à partir de l'échantillon d'apprentissage $\mathcal{D}_N$ comme suit.

Estimation des fonctions de dérive Pour chaque étiquette $i \in \mathcal{Y}$, on estime la fonction de dérive b_i^* sur $\mathbb{R}$ et par projection sur un espace contraint de dimension finie et générée par la base de fonctions **B**-splines. La construction et l'étude de la consistance des estimateurs des fonctions b_i^* sur la droite réelle $\mathbb{R}$ reposent essentiellement sur les travaux de [30] et [20]. On définit pour chaque étiquette $i \in \mathcal{Y}$, à partir du sous échantillon $\mathcal{D}_N^i = \left\{(\bar{X}^{j,i}, Y_j),\ Y_j = i,\ j = 1,\dots,N_i\right\}$ de $\mathcal{D}_N$ de taille $N_i = \sum_{j=1}^{N}\mathbb{1}_{Y_j=i}$, la fonction de contraste des moindres carrés $\gamma_{n,N}^i$ donnée pour toute fonction continue h par

$$\gamma_{n,N}^i(h) := \frac{1}{Nn}\sum_{j=1}^{N_i}\sum_{k=0}^{n-1}\left(Z_{k\Delta}^{j,i} - h(X_{k\Delta}^{j,i})\right)^2$$

où, pour tout $(j,k) \in [\![1,N_i]\!] \times [\![0,n-1]\!]$,

$$Z_{k\Delta_n}^{j,i} = \frac{X_{(k+1)\Delta_n}^{j,i} - X_{k\Delta}^{j,i}}{\Delta_n}.$$

L'estimateur non-paramétrique $\widetilde{b}_i$ de la fonction de dérive b_i^* est donné comme suit :

$$\widetilde{b}_i \in \underset{h\in\mathcal{S}_{m_i,L_i}}{\arg\min}\gamma_{n,N}^i(h)$$

où

$$\mathcal{S}_{m_i,L_i} = \left\{h = \sum_{\ell=0}^{m_i-1} a_\ell B_\ell,\ \sum_{\ell=0}^{m_i-1} a_\ell^2 \le m_i L_i\right\}$$

est l'espace contraint de dimension finie $m_i \geq 1$, généré par une base de fonctions **B**-splines $(B_\ell)_{\ell=0,\dots,m_i-1}$ construite sur un intervalle borné $[-A_i, A_i]$ avec $A_i > 0$, et $L_i > 0$ est une constante dépendant de la la classe i. Les variables m_i, L_i et A_i dépendent de N_i. L'espace $\mathcal{S}_{m_i,L_i}$ est utilisé pour approximer l'espace initial contenant b_i^*. On tronque ensuite l'estimateur $\widetilde{b}_i$ et on obtient l'estimateur tronqué $\widehat{b}_i$ donné pour tout $x \in \mathbb{R}$ par

$$\widehat{b}_i(x) = \widetilde{b}_i(x)\mathbb{1}_{|\widetilde{b}_i(x)|\leq\sqrt{L_i}} + \text{signe}(\widetilde{b}_i(x))\mathbb{1}_{|\widetilde{b}_i(x)|>\sqrt{L_i}}.$$

Estimation du coefficient de diffusion Nous considérons l'estimateur de σ^{*2} qu'on note $\widetilde{\sigma}^2$ construit dans la section 1.4.1. L'estimateur tronqué $\widehat{\sigma}^2$ est donnée pour tout $x \in \mathbb{R}$ par

$$\widehat{\sigma}^2(x) = \widetilde{\sigma}^2(x)\mathbb{1}_{\frac{1}{\log(N)}<\widetilde{\sigma}^2(x)\leq\sqrt{L}} + \sqrt{L}\mathbb{1}_{\widetilde{\sigma}^2(x)>\sqrt{L}} + \frac{1}{\log(N)}\mathbb{1}_{\widetilde{\sigma}^2(x)\leq 1/\log(N)}.$$

La consistance du classifieur *plug-in* $\widehat{g} = g_{\widehat{\mathbf{b}},\widehat{\sigma}^2,\widehat{\mathbf{p}}}$ est déduite de la consistance des estimateurs $\widehat{b}_i$ et $\widehat{\sigma}^2$ obtenue sous certaines hypothèses sur les fonctions b_i^*, σ^*, et les quantités m, m_i, L et L_i comme énoncé ci-dessous.

Théorème 1.4.8. *Supposons que*

$$\inf_{h\in\mathcal{S}_{m,L}} \|h - \sigma^{*2}\|_n^2 = \mathrm{O}\left(\frac{1}{m^2}\right), \quad \forall\, i \in \mathcal{Y}, \quad \inf_{h\in\mathcal{S}_{m_i,L_i}} \|h - b_i^*\|_n^2 = \mathrm{O}\left(\frac{1}{m_i^2}\right).$$

Sous les hypothèses 1.2.2 et 1.3.3 et sous des hypothèses sur m_i, m, L, L_i, A, A_i et N_i, l'excès de risque du classifieur plug-in $\widehat{g}$ satisfait :

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \underset{n,N\to\infty}{\longrightarrow} 0.$$

Le résultat de consistance est établi dans le cas général ou les fonctions de dérive, le coefficent de diffusion et la loi discrète de l'étiquette Y sont inconnus. Nous obtenons ensuite, et sous les mêmes hypothèses sur les fonctions inconnues b_i^* et σ^*, la vitesse de convergence énoncé ci-dessous.

Théorème 1.4.9. *Supposons que*

$$\inf_{h\in\mathcal{S}_{m,L}} \|h - \sigma^{*2}\|_n^2 = \mathrm{O}\left(\frac{1}{m^2}\right), \quad \forall\, i \in \mathcal{Y}, \quad \inf_{h\in\mathcal{S}_{m_i,L_i}} \|h - b_i^*\|_n^2 = \mathrm{O}\left(\frac{1}{m_i^2}\right).$$

Sous les hypothèses 1.2.2 et 1.3.3 et sous des hypothèses sur m_i, m, L, L_i, A, A_i et N_i, il existe des constantes $C > 0$ et $c > 0$ telles que

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\exp\left(\sqrt{c\log(N)}\right)N^{-1/5}.$$

Notons que la vitesse établie au Théorème 1.4.9 découle directement du Théorème 1.4.8 et de la vitesse de convergence des estimateurs des fonctions de dérive et du coefficient de diffusion. Nous avons établi que l'estimateur du coefficient de diffusion atteind une vitesse de convergence plus rapide par rapport aux estimateurs des fonctions de dérive. Ainsi, la vitesse atteint par le classifieur empirique $\widehat{g}$ est celui des estimateurs des fonctions de dérive $\widehat{b}_i,\ i \in \mathcal{Y}$.

Vitesse de convergence du classifieur *plug-in* $\widehat{g}$ lorsque $\sigma^* \equiv 1$. Supposons que le coefficient de diffusion σ^* est connu, constant et égal à 1. Nous pouvons ainsi établir des approximations plus fines des densités de transition des processus de diffusion $X_{|Y=i}$ avec $i \in \mathcal{Y}$. Ces approximations sont essentielles pour l'établissement de vitesses de convergence plus rapides comparées au résultat du théorème 1.4.9.

Si nous supposons en plus que les fonctions de dérive b_i^* sont bornées, nous obtenons alors le résultat suivant.

Theorem 1.4.10. *Supposons que $\sigma^* = 1$ et que les fonctions de dérive b_i^* appartiennent à la classe de Hölder $\Sigma(\beta, R)$ avec $\beta \geq 1$ le paramètre de régularité et $R > 0$. Alors, sous des hypothèses sur les fonctions b_i^* et les quantités m, m_i, L, L_i, A, A_i et Δ_n, il existe des constantes $c > 0$ et $C > 0$ telles que*

$$\mathbb{E}\left[\mathcal{R}\left(\widehat{g}\right) - \mathcal{R}\left(g^*\right)\right] \leq C \exp\left(\sqrt{c\log(N)}\right) N^{-\beta/(2\beta+1)}.$$

La vitesse obtenue (à un facteur d'ordre $\exp\left(\sqrt{c\log(N)}\right)$ près) est du même ordre que la vitesse optimale établie dans [4] sous des hypothèses de complexité sur la fonction de régression en classification binaire de données vectorielles.

Création d'un package R. Nous avons créé un package `R` appelé `SDEclassif` et disponible sur Github (lien SDEclassif). Ce package contient la fonction `SDEclassif` qui entraine un classifieur *plug-in* pour les trajectoires de diffusion, suivant la procédure de classification étudiée dans cette contribution et détaillée dans le chapitre 3.

1.4.3 Classifieur de type ERM pour les trajectoires de diffusion

Nous reconsidérons le modèle de diffusion (1.1). Nous définissons, à partir de l'échantillon d'apprentissage $\mathcal{D}_N = \left((\bar{X}^j, Y_j)\right)_{j=1,\ldots,N}$ constitué de N copies indépendante de $(\bar{X}, Y)$ avec $\bar{X} = (X_{k\Delta_n})_{0\leq k\leq n}$, le risque empirique de classification $\widehat{\mathfrak{R}}$ donné par l'équation (1.17). Ainsi, le classifieur empirique $g_{\widehat{h}}$ de type ERM est caractérisé par l'équation (1.15), où la fonction score $\widehat{h}$ est donnée par l'équation (1.16).

Nous approximons les deux espaces de fonctions contenant respectivement les fonctions de dérive $b_i^*,\ i \in \mathcal{Y}$ dune part, et le coefficient de diffusion σ^* d'autre part, par chacun des sous-espaces contraints respectifs $\mathcal{S}_m$ et $\widetilde{\mathcal{S}}_m$ de dimension $m \geq 1$ générés par une base de fonctions **B**-splines et donnés par :

$$\mathcal{S}_m := \left\{ \mathfrak{f} = \sum_{\ell=0}^{m-1} a_\ell B_\ell,\ \ \sum_{\ell=0}^{m-1} a_\ell^2 \leq m\log^3(N) \right\}, \tag{1.31}$$

et,

$$\widetilde{\mathcal{S}}_m := \left\{ x \in \mathbb{R} \mapsto \mathfrak{s}^2(x) = \mathfrak{f}(x)\mathbb{1}_{\mathfrak{f}(x)\geq 1/\log^2(N)} + \frac{1}{\log^2(N)}\mathbb{1}_{\mathfrak{f}(x)\leq 1/\log^2(N)},\ \ \mathfrak{f} \in \mathcal{S}_m \right\}. \tag{1.32}$$

La dimension m dépend de la taille N de l'échantillon d'apprentissage. La définition de l'espace d'approximation $\widetilde{\mathcal{S}}_m$ tient compte du fait que la fonction σ^{*2} est bornée inférieurement. Ainsi, nous établissons le résultat de consistance suivant.

Théorème 1.4.11. *Supposons que les fonctions b_i^* et σ^{*2} appartiennent à l'espace Hölder $\Sigma(\beta, R)$ de paramètre de régularité $\beta \geq 1$ avec $R > 0$. Sous des hypothèses sur $b_i^*,\ \sigma^{*2}$ et m, on a :*

$$\mathbb{E}\left[\mathcal{R}(g_{\widehat{h}}) - \mathcal{R}(g_{h^*})\right] = \mathrm{O}\left(N^{-\beta/(2\beta+1)}\right).$$

On remarque d'une part que cette vitesse est établie dans le cas général où les fonctions dérive b_i^*, le coefficient de diffusion σ^* et la loi $\mathfrak{p}^*$ de Y sont inconnus. Et d'autre part, le classifieur de type ERM a une vitesse de convergence plus rapide comparé au classifieur de type *plug-in*. Ce résultat s'explique par le fait que le classifieur de type ERM est directement lié à la minimisation du risque empirique de classification, contrairement au classifieur de type *plug-in* dont la construction n'est pas basée sur l'optimisation de sa performance.

Hypothèse de marge sur la fonction de régression en classification binaire. Dans ce qui suit, on se place en classification binaire, c'est à dire que $K = 2$ et $\mathcal{Y} = \{0, 1\}$. Le classifieur de Bayes g^* est donné pour $X \in \mathcal{X}$ par $g^*(X) = \mathbb{1}_{\eta^*(X)>1/2}$ avec $\eta^*(X) = \mathbb{P}(Y = 1|X)$. On suppose que $\sigma^* = 1$, les fonctions de dérive b_0^* et b_1^* sont bornées, et la variable aléatoire

$$Z := \frac{1}{\Delta_{b^*}} \int_0^1 (b_1^* - b_0^*)(X_s)dW_s, \ \text{ avec } \ \Delta_{b^*}^2 = \mathbb{E}\left[\int_0^1 (b_1^* - b_0^*)^2(X_s)ds\right] > 0,$$

admet une densité bornée sur $\mathbb{R}$. Nous établissons le premier résultat ci-dessous.

Proposition 1.4.12. *Sous des hypothèses sur b_0^* et b_1^*, et pour tout $\varepsilon \in (0, 1/8)$, il existe une constante $C > 0$ telle que*

$$\mathbb{P}\left(0 < \left|\eta^*(X) - \frac{1}{2}\right| \leq \varepsilon\right) \leq C\frac{12}{\Delta_{b^*}}\varepsilon.$$

Le résultat de la proposition 1.4.12 est connu sous l'appelation hypothèse de marge sur la fonction de régression η^* en classification binaire. Elle sert, sur le plan théorique, à minimiser les erreurs de classification commises par g^* lorsque la probabilité d'appartenance de X à la classe 1 est suffisamment proche de $1/2$. Nous obtenons ainsi le résultat suivant.

Théorème 1.4.13. *On suppose que $b_0^*,\ b_1^* \in \Sigma(\beta, R)$ avec $\beta \geq 1,\ R > 0$, et $\sigma^* = 1$. Sous des hypothèses supplémentaires sur les fonctions $b_i^*,\ i = 0, 1$ et m, on obtient :*

$$\mathbb{E}\left[\mathcal{R}(g_{\widehat{h}}) - \mathcal{R}(g_{h^*})\right] = \mathrm{O}\left(N^{-4\beta/3(2\beta+1)}\right).$$

La vitesse de convergence obtenue est plus rapide que celle établie dans un cadre plus général, et en l'absence d'hypothèse de marge sur la fonction de régression associée au modèle.

1.4.4 Illustration numérique

Dans cette section, nous évaluons numériquement les classifieurs de types *plug-in* et ERM pour les trajectoires de diffusion. Pour cette illustration, nous fixons le nombre de classes à $K = 3$, et la loi discrète de $Y \in \{1, 2, 3\}$ est $\mathfrak{p}^* = (1/3, 1/3, 1/3)$. Nous considérons le modèle de diffusion ci-dessous :

$$\text{Model : } b(\theta, x) = \theta\left[1/4 + (3/4)\cos^2 x\right],\ \ \theta \in \{1, 3/2, -3/2\},\ \ \sigma(x) = 0.1 + 0.9/\sqrt{1 + x^2}.$$

Nous évaluons le classifieur de Bayes sur des échantillons de taille $N = 4000$ contenant des trajectoires observées en temps discrets avec le pas de temps $\Delta_n = 1/n$ et $n = 500$. L'évaluation est réalisée sur 100 répétitions et nous obtenons le résultats ci-dessous.

$\widehat{\mathcal{R}}(g^*)$	0.36 (0.01)

TABLE 1.1 : *Risque de classification du classifieur de Bayes g^* à partir d'échantillons d'apprentissage de taille $N = 4000$ avec $n = 500$.*

La performance numérique du classifieur de Bayes nous permet d'évaluer la complexité du modèle de diffusion choisi. Elle permet également, comme nous l'avons mentionné en théorie, d'évaluer la performance des classifieurs empiriques construits à partir de données observées.

Nous évaluons la performance des classifieurs empiriques de types *plug-in* et ERM à partir d'échantillons d'apprentissage de taille $N \in \{100, 1000\}$. Le classifieur empirique de type *plug-in* est implémenté dans le package `SDEclassif` disponible sur github (lien SDEclassif). Le tableau ci-dessous compare les performances obtenues avec celles des classifieurs existants dans la littérature et basés sur la forêt aléatoire et la profondeur des données.

Pour le cas des classifieurs de type *plug-in* et ERM, l'espace d'approximation généré par une base de fonctions **B**-splines est choisi telle $M = 3$ et $K_N \in \mathcal{K} = \{1, 2, 4, 8, 16\}$. Pour chaque étiquette $i \in \{1, 2, 3\}$, K_{N_i} est choisi à la main pour l'évaluation du classifieur de type ERM. Pour l'évaluation du classifieur de type *plug-in* basé sur des estimateurs non-paramétriques des coefficients de dérive, on sélectionne la dimension $\widehat{K}_{N_i} + M$ telle que

$$\widehat{K}_{N_i} := \underset{K\in\mathcal{K}}{\operatorname{argmin}}\left\{\frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}(\widehat{b}_{i,K} - Z^j_{k\Delta_n})^2 + \frac{(K+M)\log^3(N)}{N}\right\}$$

Classifieurs	$N = 100$	$N = 1000$
Plug-In	0.39 (0.06)	0.37 (0.05)
ERM	0.38 (0.02)	0.36 (0.01)
Forêt aléatoire	0.40 (0.01)	0.38 (0.01)
Profondeur	0.40 (0.02)	0.39 (0.01)

TABLE 1.2 : *Performance des classifieurs de types plug-in et ERM, et comparaison avec la performance des classifieurs basés sur la forêt aléatoires et la profondeur des données, pour des trajectoires de diffusions collectées en temps discrets avec* $n = 100$.

pour l'estimation adaptative de σ^{*2}, on sélectionne la dimension $\widehat{K}_N$ telle que

$$\widehat{K}_N := \underset{K\in\mathcal{K}}{\operatorname{argmin}} \left\{ \frac{1}{Nn} \sum_{j=1}^{N} \sum_{k=0}^{n-1} (\widehat{\sigma}_K^2 - U^j_{k\Delta_n})^2 + \frac{5(K + M\log^3(N))}{Nn} \right\}.$$

Pour plus de précision sur la sélection de dimension pour l'estimation des fonctions de dérive et du coefficient de diffusion, nous renvoyons le lecteur respectivement à l'article [30] et au chapitre 2 (ou à l'article [35]).

D'après les résultats de la Table 1.2, nous observons que nos classifieurs ont une meilleure performance comparés aux deux classifieurs pour des données fonctionnelles proposés dans la littérature. Nous remarquons également que le classifieur de type ERM est numériquement plus efficace que celui de type *plug-in*. Ce résultat illustre les résultats théoriques obtenus pour chacun de ces classifieurs. Rappelons toutefois que l'évaluation numérique du classifieur de type ERM n'est pas optimale, puisque les dimensions des espaces d'approximation sont choisies à la main.

1.5 Conclusion et perspectives de recherche

Cette thèse s'est focalisée sur une étude statistique, dans un cadre non-paramétrique, des trajectoires générées par un mélange de diffusions unidimensionnelles, solutions d'une équation différentielle stochastique homogène en temps, dont le coefficient de dérive b^*_Y, supposé inconnu, dépend de la classe $Y \in \{1, \dots, K\}$ de loi discrète $\mathfrak{p}^*$ inconnue, et le coefficient de diffusion σ^*, également supposé inconnu, est commun à toutes les classes. Cette étude a abouti à l'obtention de résultats de consistance, tant pour l'estimation de σ^{*2} que pour les classifieurs de types *plug-in* et ERM.

Les prochains travaux de recherche sur l'étude statistique des trajectoires de diffusion porteront principalement sur les thématiques suivantes :

i) **Proposition d'un estimateur adaptatif du classifieur de Bayes g^* par la minimisation du risque empirique de classification.**

Nous avons, d'après l'équation (1.19) suivie de la proposition 1.3.4, le classifieur de Bayes g^* est défini comme une fonction du vecteur de coefficients de dérive $\mathbf{b}^* = (b^*_1, \dots, b^*_K)$, du coefficient de diffusion σ^* et de la loi discrète $\mathfrak{p}^*$. Autrement dit, il existe une fonction $\mathfrak{g}$ telle que :

$$g^* = \mathfrak{g}(\mathbf{b}^*, \sigma^*, \mathfrak{p}^*).$$

En convexifiant le modèle, on a $g^* = g_{h^*}$ où h^* est la fonction score qui minimise le risque défini avec la perte quadratique et donné dans l'équation (1.18). Il existe également une fonction $\mathfrak{h}$ telle que :

$$h^* = \mathfrak{h}(\mathbf{b}^*, \sigma^*, \mathfrak{p}^*).$$

Ainsi, nous proposons dans le chapitre 4 un classifieur empirique $g_{\widehat{h}}$ en minimisant le risque empirique $\widehat{\mathfrak{R}}$ donné dans (1.17) sur des espaces d'approximation des dérives b_i^* et du coefficient de diffusion σ^*. Plus précisément, on a pour tous $m, m' \geq 1$,

$$\widehat{h}_{m,m'} = \underset{(\widetilde{\mathfrak{b}},\widetilde{\sigma},\mathfrak{p})\in\mathcal{H}_{m,m'}}{\arg\min} \widehat{\mathfrak{R}}[\mathfrak{h}(\widetilde{\mathfrak{b}},\widetilde{\sigma},\mathfrak{p})]$$

avec,

$$\mathcal{H}_{m,m'} = (\mathcal{S}_m)^K \times \widetilde{\mathcal{S}}_{m'} \times (0,1)^K$$

et les espaces d'approximation $\mathcal{S}_m$ et $\widetilde{\mathcal{S}}_{m'}$ sont données respectivement par les équations (1.31) et (1.32). Les dimension des espaces d'approximation sont jusque là choisies à la main. L'objectif est d'effectuer, à partir de l'échantillon d'apprentissage, une sélection de la dimension de chaque espace d'approximation dans un ensemble fini $\mathcal{M}_N$ des valeurs possibles des dimensions, défini en fonction de la taille N de l'échantillon d'apprentissage.

ii) **Établissement de vitesses optimales.**

Un objectif serait :

- Soit de prouver l'optimalité de certaines vitesses obtenues dans ce manuscrit, citons par exemple les résultats du théorème 1.4.4 ou du corollaire 1.4.2 pour l'estimation de σ^{*2}, ou encore les théorèmes 1.4.10 et 1.4.11 pour la consistance des classifieurs de types *plug-in* et ERM.
- Soit d'étudier une vitesse optimale dans les autres cas.

iii) **Extension au cas des diffusions en grande dimension.**

Le but est de considérer les processus de diffusion $X = (X_t)_{t\in[0,T]} = \left(X_t^1, \ldots, X_t^d\right)_{t\in[0,T]}$ de dimension $d \geq 2$, et un modèle de diffusion défini comme suit :

$$dX_t = b_Y^*(X_t)dt + \Sigma^*(X_t)dW_t, \;\; t \in [0,T], \;\; X_0 = x_0 \in \mathbb{R}^d, \tag{1.33}$$

où $T \in \mathbb{R}_{+*}$ est l'horizon temporel, $(W_t)_{t\geq 0} = \left(W_t^1, \ldots, W_t^d\right)_{t\geq 0}$ est un mouvement brownien de dimension d, la fonction de dérive $b_Y^* = \left(b_Y^{*1}, \ldots, b_Y^{*d}\right)$ est supposée inconnue et dépend de l'étiquette $Y \in \{1, \ldots, K\}$, avec $K \geq 2$. On suppose que Y est toujours une variable aléatoire indépendante du mouvement brownien $(W_t)_{t\geq 0}$ et de loi discrète $\mathfrak{p}^*$ inconnue. Enfin, le coefficient de diffusion multidimensionnel $\Sigma^* = \left(\sigma_{i,j}^*\right)_{1\leq i,j\leq d}$ est commun à toutes les classes. On distinguera le cas où Σ^* est connu, puis considérer un cadre plus général avec Σ^* inconnu.

L'objection est de construire des procédures de classification non-paramétriques de type *plug-in* et de type ERM (minimisation du risque empirique). Comme les classifieurs de type *plug-in* sont basés sur des estimateurs non-paramétriques des coefficients du processus de diffusion, leur étude entrainera inéluctablement une étude sur l'estimation non-paramétrique de ces fonctions, contribuant ainsi à l'enrichissement d'une littérature déjà existante sur l'inférence statistique des processus de diffusion en grande dimension (voir par exemple [84], [76], [14]).

iv) **Évaluation des procédures de classification sur des données réelles.**

Une application sur des données réelles peut être envisagée en biologie cellulaire, les mouvements de cellules étant modélisables par des processus à sauts. Cette application nécessitera ainsi une extension de l'étude pour la classification des données de diffusion à sauts. On pourra toutefois compter sur la robustesse des procédures déjà construites, et les utiliser directement

pour l'identification de types de cellules en se basant sur leur mode de déplacement dans l'organisme.

En plus de la biologie cellulaire, nous avons des domaines d'application possibles tels que la finance, ou encore la météorologie.

1.6 Annexes

Pour simplifier les notations, nous posons $b = b^*$ et $\sigma = \sigma^*$.

Preuve de la proposition 1.2.1. On part de l'équation différentielle stochastique (1.1) rappelée ci-dessous :

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t.$$

Premier résultat. On fixe $s \in [0,1)$ et pour tout $t \in (s,1]$,

$$\frac{X_t - X_s}{t-s} = \frac{1}{t-s}\int_s^t b(X_u)du + \frac{1}{t-s}\int_s^t \sigma(X_u)dW_u$$
$$\mathbb{E}\left[\frac{X_t - X_s}{t-s}\right] = \mathbb{E}\left[\frac{1}{t-s}\int_s^t b(X_u)du\right].$$

Ainsi, on obtient de la continuité de la dérive b et du processus de diffusion X, et du théorème de convergence dominée :

$$\begin{aligned}\lim_{t\to s^+}\mathbb{E}\left[\frac{X_t - X_s}{t-s}\right] &= \lim_{t\to s^+}\mathbb{E}\left[\frac{1}{t-s}\int_s^t b(X_u)du\right] \\ &= \mathbb{E}\left[\lim_{t\to s^+}\frac{1}{t-s}\int_s^t b(X_u)du\right] \\ &= \mathbb{E}[b(X_s)].\end{aligned}$$

Deuxième résultat. On fixe $s \in [0,1)$ et pour tout $t \in (0,1]$, on a :

$$\frac{\langle X,X\rangle_t - \langle X,X\rangle_s}{t-s} = \frac{1}{t-s}\int_s^t \sigma^2(X_u).$$

Finalement, on déduit de la continuité de la fonction σ^2 et du processus de diffusion X que

$$\lim_{t\to s^+}\frac{\langle X,X\rangle_t - \langle X,X\rangle_s}{t-s} = \lim_{t\to s^+}\frac{1}{t-s}\int_s^t \sigma^2(X_u) = \sigma^2(X_s).$$

□

Preuve de la proposition 1.2.4. Soit $q \geq 1$. Supposons que $X_0 = 0$. Sous l'hypothèse 3.2.1 nous déduisons de l'équation (1.1) que pour tout $t \in [0,1]$,

$$\begin{aligned}X_t &= \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dW_s \\ |X_t| &\leq \int_0^t |b(X_s)|ds + \left|\int_0^t \sigma(X_s)dW_s\right| \\ |X_t| &\leq \int_0^t C_0(1+|X_s|)ds + \left|\int_0^t \sigma(X_s)dW_s\right|\end{aligned}$$

où la constante $C_0 > 0$ dépend de la constante $L_0 > 0$ donnée dans l'hypothèse 3.2.1. Soit $p > 1$ un réel tel que $1/p + 1/q = 1$. D'après l'inégalité de Hölder, il existe une constante C_q dépendant de q telle que

$$\forall\, t \in [0,1],\ \ |X_t|^q \leq C_q\left(\int_0^t (1+|X_s|^q)ds + \left|\int_0^t \sigma(X_s)dW_s\right|^q\right).$$

En utilisant l'inégalité de Cauchy Schwarz, on déduit que

$$\mathbb{E}\left[\sup_{t\in[0,1]}|X_t|^q\right] \leq C_q\left(\int_0^1(1+\mathbb{E}\left[|X_s|^q\right])ds+\mathbb{E}\left[\sup_{t\in[0,1]}\left|\int_0^t\sigma(X_s)dW_s\right|^q\right]\right)$$
$$\leq C_q\left(\int_0^1(1+\mathbb{E}\left[|X_s|^q\right])ds+\sqrt{\mathbb{E}\left[\sup_{t\in[0,1]}\left|\int_0^t\sigma(X_s)dW_s\right|^{2q}\right]}\right).$$

D'après l'inégalité de Doob (voir [81], corollary 1.6), on a

$$\mathbb{E}\left[\sup_{t\in[0,1]}\left|\int_0^t\sigma(X_s)dW_s\right|^q\right] \leq \left(\frac{2q}{2q-1}\right)^q\sqrt{\mathbb{E}\left[\left|\int_0^1\sigma(X_s)dW_s\right|^{2q}\right]}.$$

De plus, en utilisant l'inégalité de Burkholder-Davis-Gundy, il existe une constante $C_{1,q}$ dépendant de $q\geq 1$ telle que

$$\sqrt{\mathbb{E}\left[\left|\int_0^1\sigma(X_s)dW_s\right|^{2q}\right]} \leq C_{1,q}\mathbb{E}\left[\int_0^1\sigma^q(X_s)ds\right] < C_{1,q}\sigma_1^q.$$

On déduit donc que

$$\mathbb{E}\left[\sup_{t\in[0,1]}|X_t|^q\right] \leq C_q\left(\int_0^1(1+\mathbb{E}\left[|X_s|^q\right])ds+C_{1,q}\sigma_1^q\right).$$

Finalement, d'après la proposition 1.2.3, on a pour tout $q\geq 1$ et pour tout $s\in[0,1]$, $\mathbb{E}\left[|X_s|^q\right]<\infty$. □

***Preuve de la proposition** 1.2.6.* D'après l'équation (1.1), on a

$$dX_t = b(X_t)dt+\sigma(X_t)dW_t$$

et pour tout $k\in[\![0,n-1]\!]$,

$$\begin{aligned}
\left(X_{(k+1)\Delta_n}-X_{k\Delta_n}\right)^2 &= \left(\int_{k\Delta_n}^{(k+1)\Delta_n}b(X_s)ds+\int_{k\Delta_n}^{(k+1)\Delta_n}\sigma(X_s)dW_s\right)^2\\
&= \left(\int_{k\Delta_n}^{(k+1)\Delta_n}b(X_s)ds\right)^2+2\left(\int_{k\Delta_n}^{(k+1)\Delta_n}b(X_s)ds\right)\left(\int_{k\Delta_n}^{(k+1)\Delta_n}\sigma(X_s)dW_s\right)\\
&\quad+\left(\int_{k\Delta_n}^{(k+1)\Delta_n}\sigma(X_s)dW_s\right)^2\\
&= 2\left(\int_{k\Delta_n}^{(k+1)\Delta_n}(b(X_s)-b(X_{k\Delta_n}))ds\right)\left(\int_{k\Delta_n}^{(k+1)\Delta_n}\sigma(X_s)dW_s\right)\\
&\quad+\left(\int_{k\Delta_n}^{(k+1)\Delta_n}b(X_s)ds\right)^2+\left(\int_{k\Delta_n}^{(k+1)\Delta_n}\sigma(X_s)dW_s\right)^2-\int_{k\Delta_n}^{(k+1)\Delta_n}\sigma^2(X_s)ds\\
&\quad+\int_{k\Delta_n}^{(k+1)\Delta_n}\sigma^2(X_s)ds+2\Delta_n b(X_{k\Delta_n})\int_{k\Delta_n}^{(k+1)\Delta_n}\sigma(X_s)dW_s.
\end{aligned}$$

On déduit alors que

$$\left(X_{(k+1)\Delta_n}-X_{k\Delta_n}\right)^2 = \int_{k\Delta_n}^{(k+1)\Delta_n}\sigma^2(X_s)ds+\left(\int_{k\Delta_n}^{(k+1)\Delta_n}b(X_s)ds\right)^2+\Delta_n R^{(2)}_{k\Delta_n}+\Delta_n\zeta^{(1)}_{k\Delta_n}+\Delta_n\zeta^{(3)}_{k\Delta_n}. \tag{1.34}$$

Comme la fonction σ est de classe $\mathcal{C}^2$ sur $\mathbb{R}$, d'après la formule d'Itô, on a

$$\begin{aligned} d(\sigma^2(X_t)) &= 2\sigma'(X_t)\sigma(X_t)dX_t + 2\left[\sigma'' + (\sigma')^2\right](X_t)dt \\ &= \Phi(X_t)dt + 2\sigma'(X_t)\sigma^2(X_t)dW_t \end{aligned}$$

avec $\Phi = 2b\sigma'\sigma + \left[\sigma''\sigma + (\sigma')^2\right]\sigma^2$. On déduit que pour tous $k \in [\![0, n-1]\!]$ et $s \in [k\Delta_n, (k+1)\Delta_n)$ tels que,

$$\sigma^2(X_s) = \sigma^2(X_{k\Delta_n}) + \int_{k\Delta_n}^{s} \Phi(X_u)du + 2\int_{k\Delta_n}^{s} \sigma'(X_u)\sigma^2(X_u)dW_u.$$

On intègre ensuite sur l'intervalle $[k\Delta_n, (k+1)\Delta_n)$ et on obtient

$$\begin{aligned} \int_{k\Delta_n}^{(k+1)\Delta_n} \sigma^2(X_s)ds &= \Delta_n\sigma^2(X_{k\Delta_n}) + \int_{k\Delta_n}^{(k+1)\Delta_n}\int_{k\Delta_n}^{s} \Phi(X_u)duds \\ &\quad + 2\int_{k\Delta_n}^{(k+1)\Delta_n}\int_{k\Delta_n}^{s} \sigma'(X_u)\sigma^2(X_u)dW_uds \\ &= \Delta_n\sigma^2(X_{k\Delta_n}) + \int_{k\Delta_n}^{(k+1)\Delta_n} \Phi(X_u)\int_{u}^{(k+1)\Delta_n} dsdu \\ &\quad + 2\int_{k\Delta_n}^{(k+1)\Delta_n} \sigma'(X_u)\sigma^2(X_u)\int_{u}^{(k+1)\Delta_n} dsdW_u \\ &= \Delta_n\sigma^2(X_{k\Delta_n}) + \int_{k\Delta_n}^{(k+1)\Delta_n} [(k+1)\Delta_n - u]\Phi(X_u)du \\ &\quad + 2\int_{k\Delta_n}^{(k+1)\Delta_n} [(k+1)\Delta_n - u]\sigma'(X_u)\sigma(X_u)dW_u. \end{aligned}$$

Ainsi, on obtient pour tout $k \in [\![0, n-1]\!]$,

$$\int_{k\Delta_n}^{(k+1)\Delta_n} \sigma^2(X_s)ds = \Delta_n\sigma^2(X_{k\Delta_n}) + \int_{k\Delta_n}^{(k+1)\Delta_n} [(k+1)\Delta_n - u]\Phi(X_u)du + \zeta^{(2)}_{k\Delta_n} \tag{1.35}$$

On déduit enfin des équations (1.34) et (1.35) que

$$\frac{\left(X_{(k+1)\Delta_n} - X_{k\Delta_n}\right)^2}{\Delta_n} = \sigma^2(X_{k\Delta_n}) + \zeta_{k\Delta_n} + R_{k\Delta_n}.$$

□

Preuve de la proposition 1.3.2. Pour tout $g \in \mathcal{G}(\mathcal{X}, \mathcal{Y})$, remarquons que $\sum_{i=1}^{K}\sum_{k=1}^{K} \mathbb{1}_{g^*(X)=i}\mathbb{1}_{g(X)=k} = 1$. On obtient des équations (1.10) et (1.12) :

$$\begin{aligned} \mathcal{R}(g) - \mathcal{R}(g^*) &= \sum_{i=1}^{K}\sum_{k=1}^{K} [\mathcal{R}(g) - \mathcal{R}(g^*)]\mathbb{1}_{g^*(X)=i}\mathbb{1}_{g(X)=k} \\ &= \sum_{i=1}^{K}\sum_{k=1}^{K} \mathbb{E}\left[\sum_{j=1}^{K} \pi_j^*(X)\left(\mathbb{1}_{g^*(X)=j} - \mathbb{1}_{g(X)=j}\right)\right]\mathbb{1}_{g^*(X)=i}\mathbb{1}_{g(X)=k} \\ &= \mathbb{E}\left[\sum_{i=1}^{K}\sum_{k=1}^{K}\sum_{j=1}^{K} \pi_j^*(X)\mathbb{1}_{g^*(X)=i}\mathbb{1}_{g(X)=k}\mathbb{1}_{g^*(X)=j}\right] \\ &\quad - \mathbb{E}\left[\sum_{i=1}^{K}\sum_{k=1}^{K}\sum_{j=1}^{K} \pi_j^*(X)\mathbb{1}_{g^*(X)=i}\mathbb{1}_{g(X)=k}\mathbb{1}_{g(X)=j}\right] \\ &= \Sigma_1 - \Sigma_2. \end{aligned}$$

En échangeant les indices i et j, on obtient :

$$\Sigma_1 = \sum_{i=1}^{K}\sum_{k=1}^{K}\sum_{j=1}^{K} \pi_i^*(X)\mathbb{1}_{g^*(X)=j}\mathbb{1}_{g(X)=k}\mathbb{1}_{g^*(X)=i} = \sum_{i=1}^{K}\sum_{k=1}^{K} \pi_i^*(X)\mathbb{1}_{g(X)=k}\mathbb{1}_{g^*(X)=i}$$

puis, en échangeant les indices k et j, on a :

$$\Sigma_2 = \sum_{i=1}^{K}\sum_{k=1}^{K}\sum_{j=1}^{K} \pi_k^*(X)\mathbb{1}_{g^*(X)=i}\mathbb{1}_{g(X)=j}\mathbb{1}_{g(X)=k} = \sum_{i=1}^{K}\sum_{k=1}^{K} \pi_k^*(X)\mathbb{1}_{g^*(X)=i}\mathbb{1}_{g(X)=k}.$$

Enfin, en remarquant, par définition du classifieur de Bayes g^* et sa caractérisation donnée par (1.11), que sur l'évènement $\{g^*(X) = i\}$, on a $\pi_i^*(X) \geq \pi_k^*(X)$ pour tout $k \in \mathcal{Y}$, on déduit alors que

$$\begin{aligned}\mathcal{R}(g) - \mathcal{R}(g^*) &= \mathbb{E}\left[\sum_{i=1}^{K}\sum_{k=1}^{K} \left(\pi_i^*(X) - \pi_k^*(X)\right)\mathbb{1}_{g(X)=k}\mathbb{1}_{g^*(X)=i}\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{K}\sum_{k\neq i} \left|\pi_i^*(X) - \pi_k^*(X)\right|\mathbb{1}_{g(X)=k}\mathbb{1}_{g^*(X)=i}\right].\end{aligned}$$

□

Chapter 2

Estimation non-paramétrique du coefficient de diffusion

Résumé : Considérons le processus de diffusion $X = (X_t)_{t\in[0,1]}$ observé en temps discrets et à haute fréquence, solution d'une équation différentielle stochastique dont les coefficients de dérive et de diffusion sont supposés inconnus. Dans cet article, nous nous focalisons sur l'estimation non-paramétrique du coefficient de diffusion. Nous proposons, à partir d'observations discrètes de X, des estimateurs de type ridge du carré du coefficient de diffusion, construits par minimisation d'une fonction de contraste des moindres carrés. Nous prouvons que les estimateurs sont convergents, et nous établissons des vitesses de convergence lorsque la taille de l'échantillon de données tend vers l'infini et le pas de discrétisation de l'intervalle d'observation du processus $[0, 1]$ tend vers zéro. Les résultats théoriques sont complétés par une étude numérique sur des données synthétiques.

Mots clés. Estimation non-paramétrique, processus de diffusion, coefficient de diffusion, contraste des moindres carrés, observations répétées, estimation adaptative.

Abstract : Consider a diffusion process $X = (X_t)_{t\in[0,1]}$ observed at discrete times and high frequency, solution of a stochastic differential equation whose drift and diffusion coefficients are assumed to be unknown. In this article, we focus on the nonparametric estimation of the diffusion coefficient. We propose ridge estimators of the square of the diffusion coefficient from discrete observations of X and that are obtained by minimization of the least squares contrast. We prove that the estimators are consistent and derive rates of convergence as the number of sample paths tends to infinity, and the discretization step of the time interval $[0, 1]$ tend to zero. The theoretical results are completed with a numerical study over synthetic data.

Keywords. Nonparametric estimation, diffusion process, diffusion coefficient, least squares contrast, repeated observations, adaptive estimation.

Contents

2.1 Introduction

Let $X = (X_t)_{t\in[0,1]}$ be a one dimensional diffusion process with finite horizon time, solution of the following stochastic differential equation:

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \quad X_0 = 0 \tag{2.1}$$

where $(W_t)_{t\geq 0}$ is a standard Brownian motion. The drift function b and the diffusion coefficient σ are assumed to be unknown Lipschitz functions. We denote by $(\mathcal{F}_t)_{t\in[0,1]}$ the natural filtration of the diffusion process X. The goal of the article is to construct, from N discrete observations $\bar{X}^j = (X^j_{k\Delta_n})_{0\leq k\leq n}$, $1 \leq j \leq N$ with time step $\Delta_n = 1/n$, a nonparametric estimator of the square of the diffusion coefficient $\sigma^2(.)$. We are in the framework of high frequency data since the time step Δ_n tends to zero as n tends to infinity. Furthermore, we consider estimators of $\sigma^2(.)$ built from a single diffusion path ($N = 1$), and those built on N paths when $N \to \infty$. In this paper, we first propose a ridge estimator of $\sigma^2(.)$ on a compact interval. Secondly, we focus on a nonparametric estimation of $\sigma^2(.)$ on the real line $\mathbb{R}$. We measure the risk of any estimator $\widehat{\sigma}^2$ of the square of the diffusion coefficient σ^2 by $\mathbb{E}\left[\|\widehat{\sigma}^2 - \sigma^2\|^2_{n,N}\right]$, where $\|\widehat{\sigma}^2-\sigma^2\|^2_{n,N} := (Nn)^{-1}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\left(\widehat{\sigma}^2(X^j_{k\Delta}) - \sigma^2(X^j_{k\Delta})\right)^2$ is an empirical norm defined from the sample paths.

Related works. There is a large literature on the estimation of coefficients of diffusion processes, and we focus on the papers studying the estimation of σ^2.

Estimation of the diffusion coefficient has been considered in the parametric case (see e.g. [46], [60], [47], [15], [48], [51], [87]). In the nonparametric case, estimators of the diffusion coefficient from discrete observations are proposed under various frameworks.

First, the diffusion coefficient is constructed from one discrete observation of the diffusion process ($N = 1$) in long time ($T \to \infty$) (see e.g. [57], [22], [83], [85]), or in short time ($T = 1$) (see *e.g.* [49], [88], [56], [41], [58]). Note that in short time ($T < \infty$), only the diffusion coefficient can be estimated consistently from a single discrete path contrary to the drift function whose consistent estimation relies on repeated discrete observations of the diffusion process (see e.g. [18], [30]). For the case of short time diffusion processes (for instance $T = 1$), estimators of a time-dependent diffusion coefficients $t \mapsto \sigma^2(t)$ have been proposed. In this context, [49] built a nonparametric estimator of $t \mapsto \sigma^2(t)$ and studied its L_2 risk using wavelets methods, [88] studies the L_p risk of a kernel estimator of $\sigma^2(t)$, and [56] derived a minimax rate of convergence of order $n^{-ps/(1+2s)}$ where $s > 1$

is the smoothness parameter of the Besov space $\mathcal{B}^s_{p,\infty}([0,1])$ (see later in the paper). For the space-dependent diffusion coefficient $x \mapsto \sigma^2(x)$, a first estimator based on kernels and built from a single discrete observation of the diffusion process with $T = 1$ is proposed in [42]. The estimator has been proved to be consistent under a condition on the bandwidth, but a rate of convergence of its risk of estimation has not been established.

Secondly, the diffusion coefficient is built in short time ($T < \infty$) from N repeated discrete observations with $N \to \infty$. In [28], a nonparametric estimator of σ^2 is proposed from repeated discrete observations on the real line $\mathbb{R}$ when the time horizon $T = 1$. The estimator has been proved to be consistent with a rate of order $N^{-1/5}$ over the space of Lipschitz functions.

Two main methods are used to build consistent nonparametric estimators of $x \mapsto \sigma^2(x)$. The first method is the one using kernels (see e.g. [41], [6], [80], [53], [85], [77]), the other method consists in estimating σ^2 as solution of a nonparametric regression model using the least squares approach. Since the diffusion coefficient is assumed to belong to an infinite dimensional space, the method consists in projecting σ^2 into a finite dimensional subspace, estimating the projection and making a data-driven selection of the dimension by minimizing a penalized least squares contrast (see e.g. [58], [57], [83], [17], [85], [28]).

Main contribution. In this article, we assume to have at our disposal N i.i.d. discrete observations of length n of the diffusion process X. The main objectives of this paper are the following.

1. Construct a consistent and implementable ridge estimator of σ^2 from a single diffusion path ($N = 1$) using the least squares approach. We derive rates of convergence of the risk of estimation of the ridge estimators built on a compact interval and on the real line $\mathbb{R}$ over a Hölder space, taking advantage of the properties of the local time of the diffusion process, and its link with the transition density.

2. We extend the result to the estimation of σ^2 on repeated observations of the diffusion process ($N \to \infty$). We prove that the estimators built on a compact interval and on $\mathbb{R}$ are more efficient considering their respective rates compared to nonparametric estimators built from a single diffusion path.

3. Focusing on the support of the diffusion coefficient, we consider an intermediate case between a compact interval and $\mathbb{R}$ by proposing a ridge estimator of σ^2 restricted to the compact interval $[-A_N, A_N]$ where $A_N \to \infty$ as $N \to \infty$. The benefit of this approach is that the resulting projection estimator can reach a faster rate of convergence compared to the rate obtained on the real line $\mathbb{R}$.

4. Finally, we propose adaptive estimators of σ^2 based on a data-driven selection of the dimension through the minimization of the penalized least squares contrast in different settings.

We sum up below the rates of convergence (up to a log-factor) of the ridge estimators of $\sigma^2_{|I}$ with $I \subseteq \mathbb{R}$ over a Hölder space defined in the next section with a smoothness parameter $\beta \geq 1$.

Outline of the paper. In Section 2.2, we define our framework with the key assumptions on the coefficients of the diffusion process ensuring for instance that Equation (3.1) admits a unique strong solution. Section 2.3 is devoted to the non-adaptive estimation of the diffusion coefficient from one diffusion path both on a compact interval and on the real line $\mathbb{R}$. In Section 2.4, we extend the study to the non-adaptive estimation of the diffusion coefficient from repeated observations of the diffusion process. We propose in Section 2.5, adaptive estimators of the diffusion coefficient, and Section 2.6 complete the study with numerical evaluation of the performance of estimators. We prove our theoretical results in Section 2.8.

Estimation interval	$N = 1$ and $n \to +\infty$	$N \to +\infty$ and $n \to +\infty$
$I = [-A, A],\ \ A > 0$	$n^{-\beta/(2\beta+1)}$	$(Nn)^{-\beta/(2\beta+1)}$
$I = [-A_N, A_N],\ \ A_N \underset{N\to+\infty}{\longrightarrow} +\infty$	xxxx	$(Nn)^{-\beta/(2\beta+1)},\ \ N \propto n$
$I = \mathbb{R}$	$n^{-\beta/(4\beta+1)}$	$(Nn)^{-\beta/(4\beta+1)} + n^{-2}$

Table 2.1: Rates of convergence of the square root of the risk of estimation $\mathbb{E}\left[\left\|\widehat{\sigma}^2 - \sigma^2_{|I}\right\|^2_{n,N}\right]$ of the non-adaptive estimators $\widehat{\sigma}^2$ of the square of the diffusion coefficient $\sigma^2_{|I}$ built from one diffusion path ($N = 1$) on the left column, and from repeated observations of the diffusion process ($N \to \infty$) on the right column. For the precise results, see Sections 2.3 and 2.4.

2.2 Framework and assumptions

Consider a diffusion process $X = (X_t)_{t\in[0,1]}$, solution of Equation (3.1) whose drift and diffusion coefficient satisfy the following assumption.

Assumption 2.2.1. *1. There exists a constant $L_0 > 0$ such that b and σ are L_0-Lipschitz functions on $\mathbb{R}$.*

2. There exist constants $\sigma_0, \sigma_1 > 0$ such that : $\sigma_0 \le \sigma(x) \le \sigma_1,\ \ \forall x \in \mathbb{R}$.

3. $\sigma \in \mathcal{C}^2(\mathbb{R})$ and there exist $C > 0$ and $\alpha \ge 0$ such that:

$$\left|\sigma'(x)\right| + \left|\sigma''(x)\right| \le C\left(1 + |x|^\alpha\right),\ \ \forall x \in \mathbb{R}.$$

Under Assumption 2.2.1, $X = (X_t)_{t\in[0,1]}$ is the unique strong solution of Equation (3.1), and this unique solution admits a transition density $(t, x) \mapsto p_X(t, x)$. Besides, we draw from Assumption 2.2.1 that

$$\forall q \ge 1,\ \ \mathbb{E}\left[\sup_{t\in[0,1]} |X_t|^q\right] < \infty. \tag{2.2}$$

2.2.1 Definitions and notations

We suppose to have at our disposal, a sample $D_{N,n} = \left\{\bar{X}^j,\ j = 1, \cdots, N\right\}$ constituted of N independent copies of the discrete observation $\bar{X} = (X_{k\Delta_n})_{0\le k\le n}$ of the diffusion process X where $\Delta_n = 1/n$ is the time-step. The objective is to construct, from the sample $D_{N,n}$, a nonparametric estimator of the square σ^2 of the diffusion coefficient on an interval $I \subseteq \mathbb{R}$. In the sequel, we consider two main cases, the first one being the estimation of σ^2 on the interval I from a single path ($N = 1$ and $n \to \infty$). For the second case, we assume that both N and n tend to infinity.

Denote by $\mathbb{P}_X$ the distribution of the diffusion process X, and $\mathbb{E}_X$ the corresponding expectation. For each measurable function h, such that $\mathbb{E}\left[h^2(X_t)\right] < \infty$ for all $t \in [0, 1]$, we define the following empirical norms:

$$\|h\|^2_n := \mathbb{E}_X\left[\frac{1}{n}\sum_{k=0}^{n-1} h^2\left(X_{k\Delta_n}\right)\right],\quad \|h\|^2_{n,N} := \frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1} h^2\left(X^j_{k\Delta_n}\right). \tag{2.3}$$

For all $h \in \mathbb{L}^2(I)$, we have

$$\|h\|^2_n = \int_I h^2(x)\frac{1}{n}\sum_{k=0}^{n-1} p_X(k\Delta_n, x)dx = \int_I h^2(x) f_n(x)dx,$$

where $f_n : x \mapsto \frac{1}{n}\sum_{k=0}^{n-1} p_X(k\Delta_n, x)$ is a density function. For the case of non-adaptive estimators of σ^2, we also establish bounds of the risks of the estimators based on the empirical norm $\|.\|_n$ or the $\mathbb{L}^2-$norm $\|.\|$ when the estimation interval I is compact.

For any integers $p, q \ge 2$ and any matrix $M \in \mathbb{R}^{p\times q}$, we denote by ${}^tM$, the transpose of M. In the sequel, for all $n, p \in \mathbb{N}$ such that $n > p$, we denote by $[\![p, n]\!]$, the set of integers $\{p, p + 1, \ldots n\}$.

2.2.2 Spaces of approximation

We propose projection estimators of σ^2 on a finite-dimensional subspace. To this end, we consider for each $m \geq 1$, a m-dimensional subspace $\mathcal{S}_m$ given as follows:

$$\mathcal{S}_m := \text{Span}\left(\phi_\ell,\ \ell = 0, \cdots, m-1\right),\ \ m \geq 1 \tag{2.4}$$

where the functions $(\phi_\ell,\ \ell \in \mathbb{N})$ are continuous, linearly independent and bounded on I. Furthermore, we need to control the ℓ^2-norm of the coordinate vectors of elements of $\mathcal{S}_m$, which leads to the following constrained subspace,

$$\mathcal{S}_{m,L} := \left\{ h = \sum_{\ell=0}^{m-1} a_\ell \phi_\ell,\ \ \sum_{\ell=0}^{m-1} a_\ell^2 = \|\mathbf{a}\|_2^2 \leq mL,\ \ \mathbf{a} = (a_0, \cdots, a_{m-1}),\ \ L > 0 \right\}. \tag{2.5}$$

Note that $\mathcal{S}_{m,L} \subset \mathcal{S}_m$ and $\mathcal{S}_{m,L}$ is no longer a vector space. The control of the coordinate vectors allows to establish an upper bound of the estimation error that tends to zero as $n \to \infty$ or $N, n \to \infty$. In fact, we prove in the next sections that the construction of consistent estimators of σ^2 requires the functions $h = \sum_{\ell=0}^{m-1} a_\ell \phi_\ell$ to be bounded, such that

$$\|h\|_\infty \leq \max_{\ell=0,\ldots,m-1} \|\phi_\ell\|_\infty \, \|\mathbf{a}\|_2.$$

This condition is satisfied for the functions of the constrained subspaces $\mathcal{S}_{m,L}$ with $m \geq 1$. In this article, we work with the following bases.

[B] The B-spline basis This is an example of a non-orthonormal basis defined on a compact interval. Let $A > 0$ be a real number, and suppose (without restriction) that $I = [-A, A]$. Let $K, M \in \mathbb{N}^*$, and consider $\mathbf{u} = (u_{-M}, \cdots, u_{K+M})$ a knots vector such that $u_{-M} = \cdots = u_{-1} = u_0 = -A$, $u_{K+1} = \cdots = u_{K+M} = A$, and for all $i = 0, \cdots, K$,

$$u_i = -A + i\frac{2A}{K}.$$

One calls **B**-spline functions, the piecewise polynomial functions $(B_\ell)_{\ell=-M,\cdots,K-1}$ of degree M, associated with the knots vector $\mathbf{u}$ (see [54], *Chapter 14*). The **B**-spline functions are linearly independent smooth functions returning zero for all $x \notin [-A, A]$, and satisfying some smoothness conditions established in [54]. Thus, we consider approximation subspaces $\mathcal{S}_{K+M}$ defined by

$$\mathcal{S}_{K+M} = \text{Span}\left\{B_\ell,\ \ell = -M, \cdots, K-1\right\}$$

of dimension $\dim(\mathcal{S}_{K+M}) = K + M$, and in which, each function $h = \sum_{\ell=-M}^{K-1} a_\ell B_\ell$ is $M-1$ times continuously differentiable thanks to the properties of the spline functions (see [54]). Besides, the spline basis is included in the definition of both the subspace $\mathcal{S}_m$ and the constrained subspace $\mathcal{S}_{m,L}$ (see Equations (2.4) and (2.5)) with $m = K + M$ and for any coordinates vector $(a_{-M}, \ldots, a_{K-1}) \in \mathbb{R}^{K+M}$,

$$\sum_{\ell=-M}^{K-1} a_\ell B_\ell = \sum_{\ell=0}^{m-1} a_{\ell-M} B_{\ell-M}.$$

The integer $M \in \mathbb{N}^*$ is fixed, while K varies in the set of integers $\mathbb{N}^*$. If we assume that σ^2 belongs to the Hölder space $\Sigma_I(\beta, R)$ given as follows:

$$\Sigma_I(\beta, R) := \left\{ h \in \mathcal{C}^{\lfloor\beta\rfloor+1}(I),\ \left|h^{(\ell)}(x) - h^{(\ell)}(y)\right| \leq R|x-y|^{\beta-\ell},\ \ x, y \in I \right\},$$

where $\beta \geq 1$, $\ell = \lfloor\beta\rfloor$ is the largest integer strictly smaller than β, and $R > 0$, then the unknown function $\sigma^2_{|I}$ restricted to the compact interval I can be approximated in the constrained subspace $\mathcal{S}_{K+M,L}$ spanned by the spline basis. This approximation result to the following bias term:

$$\inf_{h \in \mathcal{S}_{K+M,L}} \|h - \sigma^2_{|I}\|_n^2 \leq C|I|^{2\beta} K^{-2\beta} \tag{2.6}$$

where the constant $C > 0$ depends on β, R and M, and $|I| = \sup I - \inf I$. The above result is a modification of *Lemma D.2* in [30].

[F] The Fourier basis The subspace $\mathcal{S}_m$ can be spanned by the Fourier basis

$$\{f_\ell,\ \ell = 0,\dots,m-1\} = \{1, \sqrt{2}\cos(2\pi jx), \sqrt{2}\sin(2\pi jx),\ j = 1,...,d\} \text{ with } m = 2d+1.$$

The above Fourier basis is defined on the compact interval $[0,1]$. The definition can be extended to any compact interval, replacing the bases functions $x \mapsto f_\ell(x)$ by $x \mapsto 1/(\max I - \min I) f_\ell(\frac{x-\min I}{\max I - \min I})$. We use this basis to build the estimators of σ^2 on a compact interval $I \subset \mathbb{R}$.

Define for all $s \geq 1$ and for any compact interval $I \subset \mathbb{R}$, the Besov space $\mathcal{B}^s_{2,\infty}(I)$ which is a space of functions $f \in L^2(I)$ such that the $\lfloor s \rfloor^{th}$ derivative $f^{(\lfloor s \rfloor)}$ belongs to the space $\mathcal{B}^{s-\lfloor s \rfloor}_{2,\infty}(I)$ given by

$$\mathcal{B}^{s-\lfloor s \rfloor}_{2,\infty}(I) = \left\{ f \in L^2(I) \text{ and } \frac{w_{2,f}(t)}{t^{s-\lfloor s \rfloor}} \in L^\infty(I \cap \mathbb{R}^+) \right\}$$

where for $s - \lfloor s \rfloor \in (0,1)$, $w_{2,f}(t) = \sup_{|h| \leq t} \|\tau_h f - f\|_2$ with $\tau_h f(x) = f(x-h)$, and for $s - \lfloor s \rfloor = 1$, $w_{2,f}(t) = \sup_{|h| \leq t} \|\tau_h f + \tau_{-h} f - 2f\|_2$. Thus, if we assume that the function $\sigma^2_{|I}$ belongs to the Besov space $\mathcal{B}^s_{2,\infty}$, then it can be approximated in a constrained subspace $\mathcal{S}_{m,L}$ spanned by the Fourier basis. Moreover, under Assumption 2.2.1 and from *Lemma 12* in [2], there exists a constant $C > 0$ depending on the constant τ_1 of Equation (2.13), the smoothness parameter s of the Besov space such that

$$\inf_{h \in \mathcal{S}_{m,L}} \left\| h - \sigma^2_{|I} \right\|_n^2 \leq \tau_1 \inf_{h \in \mathcal{S}_{m,L}} \left\| h - \sigma^2_{|I} \right\|^2 \leq C \left| \sigma^2_{|I} \right|^2_\beta m^{-2\beta} \tag{2.7}$$

where $|\sigma^2_{|I}|_s$ is the semi-norm of $\sigma^2_{|I}$ in the Besov space $\mathcal{B}^s_{2,\infty}(I)$.

Note that for all $\beta \geq 1$, the Hölder space $\Sigma_I(\beta, R)$ and the Besov space $\mathcal{B}^\beta_{2,\infty}$ satisfy:

$$L^\infty(\mathbb{R}) \cap \Sigma_I(\beta, R) \subset \mathcal{B}^\beta_{\infty,\infty}(I) \subset \mathcal{B}^\beta_{2,\infty}(I)$$

(see [31], *Chap. 2 page 16*). As a result, we rather consider in the sequel the Hölder space $\Sigma_I(\beta, R)$ which can also be approximated by the Fourier basis.

[H] The Hermite basis The basis is defined from the Hermite functions $(h_j, j \geq 0)$ defined on $\mathbb{R}$ and given for all $j \geq 0$ and for all $x \in \mathbb{R}$ by:

$$h_j(x) = c_j H_j(x), \text{ where } H_j(x) = (-1)^j \exp\left(\frac{x^2}{2}\right) \frac{d^j}{dx^j}\left(\mathrm{e}^{-x^2/2}\right) \text{ and } c_j = \left(2^j j! \sqrt{\pi}\right)^{-1/2}.$$

The polynomials $H_j(x)$, $j \geq 0$ are the Hermite polynomials, and $(h_j, j \geq 0)$ is an orthonormal basis of $L^2(\mathbb{R})$. Furthermore, for all $j \geq 1$ and for all $x \in \mathbb{R}$ such that $x^2 \geq (3/2)(4j+3)$, we have $|h_j(x)| \leq c|x| \exp(-c_0 x^2)$, where $c, c_0 > 0$ are constants independent of j (see [19], Proof of Proposition 3.5). We use the Hermite basis in the sequel for the estimation of σ^2 on the real line $\mathbb{R}$.

If one assumes that σ^2 belongs to the Sobolev space $W^s_{f_n}(\mathbb{R}, R)$ given for all $s \geq 1$ by

$$W^s_{f_n}(\mathbb{R}, R) := \left\{ g \in L^2(\mathbb{R}, f_n(x)dx),\ \forall\, \ell \geq 1,\ \|g - g_\ell\|_n^2 \leq R\ell^{-s} \right\}$$

where for each $\ell \geq 1$, g_ℓ is the $L^2(\mathbb{R}, f_n(x)dx)-$orthogonal projection of g on the $\ell-$dimensional vector space $\mathcal{S}_\ell$ spanned by the Hermite basis. Consider a compact interval $I \subset \mathbb{R}$ and the following spaces:

$$W^s(I, R) := \left\{ g \in L^2(I),\ \sum_{j=0}^\infty j^s \langle g, \phi_j \rangle^2 \leq R \right\},$$
$$W^s_{f_n}(I, R) := \left\{ g \in L^2(I, f_n(x)dx),\ \forall\, \ell \geq 1,\ \|g - g_\ell\|_n^2 \leq R\ell^{-s} \right\}$$

where $(\phi_j)_{j \geq 0}$ is an orthonormal basis defined on I and for all $\ell \geq 1$, g_ℓ is the orthogonal projection of g onto $\mathcal{S}_\ell = \mathrm{Span}(h_j,\ j \leq \ell)$ of dimension $\ell \geq 1$ (see *e.g.* [20]). Then, for all $g \in W^s(I, R)$, we have

$$g = \sum_{j=0}^\infty \langle g, \phi_j \rangle \phi_j \text{ and } \|g - g_\ell\|^2 = \sum_{j=\ell+1}^\infty \langle g, \phi_j \rangle^2 \leq \ell^{-s} \sum_{j=\ell+1}^\infty j^s \langle g, \phi_j \rangle^2 \leq R\ell^{-s}.$$

We have $W^s_{f_n}(I,R) = W^s(I,R)$ as the empirical norm $\|.\|_n$ and the L^2-norm $\|.\|$ are equivalent. The space $W^s_{f_n}(\mathbb{R},R)$ is an extension of the space $W^s_{f_n}(I,R)$ wher $I=\mathbb{R}$ and $(\phi_j)_{j\geq 0}$ is the Hermite basis.

Remark 2.2.2. *The **B**-spline basis is used for the estimation of σ^2 on a compact interval on one side ($N=1$ and $N>1$), and on the real line on the other side restricting σ^2 on the compact interval $[-\log(n),\log(n)]$ for $N=1$, or $[-\log(N),\log(N)]$ for $N>1$, and bounding the exit probability of the process X from the interval $[-\log(N),\log(N)]$ (or $[-\log(n),\log(n)]$) by a negligible term with respect to the estimation error. In a similar context, the Fourier basis is used as an othonormal basis to built nonparametric estimators of σ^2 on a compact interval and on $\mathbb{R}$, both for $N=1$ and for $N>1$. The main goal is to show that, in addition to the spline basis which is not orthogonal, we can built projection estimators of σ^2 on orthonormal bases that are consistent. The advantage of the Hermite basis compared to the Fourier basis is its definition on the real line $\mathbb{R}$. As a result, we use the Hermite basis to propose for $N>1$, a projection estimator of σ^2 whose support is the real line $\mathbb{R}$.*

Remark 2.2.3. *Denote by $\mathcal{M}$, the set of possible values of the dimension $m\geq 1$ of the approximation subspace $\mathcal{S}_m$. If $(\phi_0,\cdots,\phi_{m-1})$ is an orthonormal basis, then for all $m,m'\in\mathcal{M}$ such that $m<m'$, we have $\mathcal{S}_m\subset\mathcal{S}_{m'}$. For the case of the **B**-spline basis, one can find a subset $\mathcal{K}\subset\mathcal{M}$ of the form*

$$\mathcal{K}=\{2^q,\ q=0,\cdots,q_{\max}\}$$

such that for all $K,K'\in\mathcal{K}$, $K<K'$ implies $\mathcal{S}_{K+M}\subset\mathcal{S}_{K'+M}$ (see for example [30]). The nesting of subspaces $\mathcal{S}_m,\ m\in\mathcal{M}$ is of great importance in the context of adaptive estimation of the diffusion coefficient and the establishment of upper-bounds for the risk of adaptive estimators.

In the sequel, we denote by [**F**], [**H**] and [**B**] the respective collection of subspaces spanned by the Fourier basis, the Hermite basis and the **B**-spline basis.

2.2.3 Ridge estimators of the square of the diffusion coefficient

We establish from Equation (1.1) and the sample $D_{N,n}$ the regression model for the estimation of σ^2. For all $j\in[\![1,N]\!]$ and $k\in[\![0,n-1]\!]$, define

$$U^j_{k\Delta_n}:=\frac{\left(X^j_{(k+1)\Delta_n}-X^j_{k\Delta_n}\right)^2}{\Delta_n}.$$

The increments $U^j_{k\Delta_n}$ are approximations in discrete times of $\frac{d\langle X,X\rangle_t}{dt}$ since, from Equation (1.1), one has $d\langle X,X\rangle_t=\sigma^2(X_t)dt$. From Equation (1.1), we obtain the following regression model,

$$U^j_{k\Delta_n}=\sigma^2(X^j_{k\Delta_n})+\zeta^j_{k\Delta_n}+R^j_{k\Delta_n},\ \ \forall(j,k)\in[\![1,N]\!]\times[\![0,n-1]\!] \tag{2.8}$$

where $U^j_{k\Delta_n}$ is the response variable, $\zeta^j_{k\Delta_n}$ and $R^j_{k\Delta_n}$ are respectively the error term and a negligible residual whose explicit formulas are given in Section 2.8.

We consider the least squares contrast $\gamma_{n,N}$ defined for all $m\in\mathcal{M}$ and for all functions $h\in\mathcal{S}_{m,L}$ by

$$\gamma_{n,N}(h):=\frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\left(U^j_{k\Delta_n}-h(X^j_{k\Delta_n})\right)^2. \tag{2.9}$$

For each dimension $m\in\mathcal{M}$, the projection estimator $\widehat{\sigma}^2_m$ of σ^2 over the subspace $\mathcal{S}_{m,L}$ is defined as:

$$\widehat{\sigma}^2_m\in\underset{h\in\mathcal{S}_{m,L}}{\arg\min}\,\gamma_{n,N}(h). \tag{2.10}$$

Indeed, for each dimension $m\in\mathcal{M}$, the estimator $\widehat{\sigma}^2_m$ of σ^2 given in Equation (2.10) satisfies $\widehat{\sigma}^2_m=\sum_{\ell=0}^{m-1}\widehat{a}_\ell\phi_\ell$, where

$$\widehat{\mathbf{a}}=(\widehat{a}_0,\cdots,\widehat{a}_{m-1}):=\underset{\|\mathbf{a}\|_2^2\leq mL}{\arg\min}\|\mathbf{U}-\mathbf{F}_m\mathbf{a}\|_2^2 \tag{2.11}$$

with ${}^t\mathbf{U} = \left(U_0^1, \ldots, U_{(n-1)\Delta_n}^1, \ldots, U_0^N, \ldots, U_{(n-1)\Delta_n}^N\right)$ and the matrix $\mathbf{F}_m$ is defined as follows

$$\mathbf{F}_m := \left({}^t(\phi_\ell(X_0^j), \ldots, \phi_\ell(X_{(n-1)\Delta_n}^j))\right)_{\substack{0\leq\ell\leq m-1\\ 1\leq j\leq N}} \in \mathbb{R}^{Nn\times m}.$$

The vector of coefficients $\widehat{\mathbf{a}}$ is unique and called the ridge estimator of $\mathbf{a}$ because of the ℓ^2 constraint on the coordinate vectors (see [55] Chap. 3 page 61).

2.3 Estimation of the diffusion coefficient from a single diffusion path

This section focuses on the nonparametric estimation of the square of the diffusion coefficient σ^2 on an interval $I \subseteq \mathbb{R}$ when only a single diffusion path is observed at discrete times ($N = 1$). It is proved in the literature that one can construct consistent estimators of the diffusion coefficient from one path when the time horizon T is finite (see *e.g.* [58]). Two cases are considered. First, we propose a ridge estimator of σ^2 on a compact interval $I \subset \mathbb{R}$, say for example $I = [-1, 1]$. Secondly, we extend the study to the estimation of σ^2 on the real line $I = \mathbb{R}$. Note that for case of a compact interval I, only data that fall into I are used to estimate the unknown function.

2.3.1 Non-adaptive estimation of the diffusion coefficient on a compact interval

In this section, we consider the estimator $\widehat{\sigma}_m^2$ of the compactly supported square of the diffusion coefficient $\sigma_{|I}^2$ on the constrained subspaces $\mathcal{S}_{m,L}$ from the observation of a single diffusion path.

Since the interval $I \subset \mathbb{R}$ is compact, the immediate benefit is that the density function f_n defined from the transition density of the diffusion process $\bar{X} = (X_{k\Delta})$ is bounded from below. In fact, there exist constants $\tau_0, \tau_1 \in (0, 1]$ such that

$$\forall x \in I, \quad \tau_0 \leq f_n(x) \leq \tau_1, \tag{2.12}$$

(see [30]). Thus, for each function $h \in \mathbb{L}^2(I)$,

$$\tau_0\|h\|^2 \leq \|h\|_n^2 \leq \tau_1\|h\|^2 \tag{2.13}$$

where $\|.\|$ is the $\mathbb{L}^2-$norm. Equation (2.13) allows to establish global rates of convergence of the risk of the ridge estimators $\widehat{\sigma}_m^2$ of $\sigma_{|I}^2$ with $m \in \mathcal{M}$ using the L^2-norm $\|.\|$ which is, in this case, equivalent with the empirical norm $\|.\|_n$.

Let $\|.\|_n$ be a random pseudo-norm defined for all $h \in \mathbb{L}^2(I)$ by

$$\|h\|_X^2 := \int_0^1 h^2(X_s)ds. \tag{2.14}$$

To establish an upper-bound of the risk of estimation that tends to zero as n tends to infinity, we need to establish equivalence relations between the pseudo-norms $\|.\|_{n,1}$ ($N = 1$) and $\|.\|_X$ on one side, and $\|.\|_X$ and the L^2-norm $\|.\|$ on the other side. Define for $x \in \mathbb{R}$, the local time $\mathcal{L}^x$ of the diffusion process $X = (X_t)_{t\in[0,1]}$ by

$$\mathcal{L}^x = \lim_{\varepsilon\to 0}\frac{1}{2\varepsilon}\int_0^1 \mathbb{1}_{(x-\varepsilon,x+\varepsilon)}(X_s)ds. \tag{2.15}$$

In general, the local time of a continuous semimartingale is a.s. *càdlàg* (see *e.g.* [81]). But, for diffusion processes and under Assumption 2.2.1, the local time $\mathcal{L}^x$ is continuous at any point $x \in \mathbb{R}$ (see Lemma 2.8.5 in Section 2.8). Furthermore, we obtain the following result.

Lemma 2.3.1 ([81]). *Under Assumption 2.2.1, and for any continuous and integrable function h, it holds,*

1. $\int_0^1 h(X_s)ds = \int_{\mathbb{R}} h(x)\mathcal{L}^x dx$.
2. *For all* $x \in \mathbb{R}$, $\mathbb{E}(\mathcal{L}^x) = \int_0^1 p_X(s, x)ds$.

Lemma 2.3.1 shows that there is a link between the local time and the transition density of the diffusion process. Thus, if we consider the pseudo-norm $\|.\|_X$ depending on the process $X = (X_t)_{t\in[0,1]}$ and given in Equation (2.14), and using Lemma 2.3.1, we obtain that,

$$\mathbb{E}\left[\|h\|_X^2\right] = \int_{\mathbb{R}} h^2(x)\mathbb{E}\left[\mathcal{L}^x\right]dx = \int_{\mathbb{R}} h^2(x)\int_0^1 p_X(s,x)dsdx \geq \tau_0\|h\|^2. \tag{2.16}$$

where $\int_0^1 p_X(s,x)ds \geq \tau_0 > 0$ (see [30], *Lemma 4.3*), and $\|h\|^2$ is the $\mathbb{L}^2-$norm of h.

Theorem 2.3.2. *Set $L = \log(n)$ with n large enough. Suppose that σ^2 is approximated in one of the collections* [**B**] *and* [**F**]. *Under Assumption 2.2.1 and for all $\gamma > 1$, it holds*

$$\mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_{n,1}^2\right] \leq 3\inf_{h\in\mathcal{S}_{m,L}}\left\|h - \sigma_{|I}^2\right\|_n^2 + C\left(\frac{m}{n} + \frac{m^{2\gamma+1}\log(n)}{n^{\gamma/2}} + \Delta_n^2\right)$$
$$\mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_n^2\right] \leq \frac{34\tau_1}{\tau_0}\inf_{h\in\mathcal{S}_{m,L}}\left\|h - \sigma_{|I}^2\right\|_n^2 + C'\left(\frac{m}{n} + \frac{m^{2\gamma+1}\log(n)}{n^{\gamma/2}} + \Delta_n^2\right)$$

where the constant $C > 0$ depends on σ_1 and the constant $C' > 0$ depends on σ_1, τ_0 and τ_1.

We remark that the risk bound of $\widehat{\sigma}_m^2$ is composed of the bias term, which quantifies the cost of approximation of $\sigma_{|I}^2$ in the constrained space $\mathcal{S}_{m,L}$, the estimation error $\mathrm{O}(m/n)$ and the cost of the time discretization $\mathrm{O}(\Delta_n^2)$ are established on a random event in which the pseudo-norms $\|.\|_{n,1}$ and $\|.\|_X$ are equivalent, and whose probability of the complementary times $\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_\infty^2$ is bounded by the term $\mathrm{O}\left(\frac{m^{2\gamma+1}\log(n)}{n^{\gamma/2}}\right)$ (see Lemma 2.8.9 and proof of Theorem 2.3.2).

The next result proves that the risk of estimation can reach a rate of convergence of the same order as the rate established in [58] if the parameter $\gamma > 1$ is chosen such that the term $\mathrm{O}(m^{2\gamma+1}\log(n)/n^{\gamma/2})$ is of the same order than the estimation error of order m/n. Note that the risk $\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_n^2$ is random since

$$\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_n^2 = \mathbb{E}_X\left[\frac{1}{n}\sum_{k=0}^{n-1}(\widehat{\sigma}_m^2 - \sigma_{|I}^2)(X_{k\Delta})\right]$$

and the estimator $\widehat{\sigma}_m^2$ is built from an independent copy $\bar{X}^1$ of the discrete times process $\bar{X}$. Thus, the expectation $\mathbb{E}$ relates to the estimator $\widehat{\sigma}_m^2$.

Corollary 2.3.3. *Suppose that $\sigma^2 \in \Sigma_I(\beta, R)$ with $\beta > 3/2$, and $\gamma = 4(2\beta+1)/(2\beta-3)$. Assume that $K_{\mathrm{opt}} \propto n^{1/(2\beta+1)}$ for* [**B**] *($m_{\mathrm{opt}} = K_{\mathrm{opt}} + M$), and $m_{\mathrm{opt}} \propto n^{1/(2\beta+1)}$ for* [**F**]. *Under Assumptions 2.2.1 and for n large enough, it holds,*

$$\mathbb{E}\left[\left\|\widehat{\sigma}_{m_{\mathrm{opt}}}^2 - \sigma_{|I}^2\right\|_{n,1}^2\right] = \mathrm{O}\left(n^{-2\beta/(2\beta+1)}\right)$$
$$\mathbb{E}\left[\left\|\widehat{\sigma}_{m_{\mathrm{opt}}}^2 - \sigma_{|I}^2\right\|_n^2\right] = \mathrm{O}\left(n^{-2\beta/(2\beta+1)}\right).$$

Note that we obtain the exact same rates when considering the risk of $\widehat{\sigma}_{m_{\mathrm{opt}}}^2$ defined with the $\mathbb{L}^2-$norm equivalent to the empirical norm $\|.\|_n$. Moreover, these rates of convergence are of the same order than the optimal rate $n^{-s/(2s+1)}$ established in [58] over a Besov ball.

2.3.2 Non-adaptive estimation of the diffusion coefficient on the real line

In this section, we propose a ridge estimator of σ^2 on the real line $\mathbb{R}$, built from one diffusion path. In this context, the main drawback is that the density function $f_n : x \mapsto \frac{1}{n}\sum_{k=0}^{n-1} p_X(k\Delta, x)$ is no longer lower bounded. Consequently, the empirical norm $\|.\|_n$ is no longer equivalent to the L_2-norm $\|.\|$ and the consistency of the estimation error is no longer ensured under the assumptions made in the previous sections. Consider the truncated estimator $\widehat{\sigma}_{m,L}^2$ of σ^2 given by

$$\widehat{\sigma}_{m,L}^2(x) = \widehat{\sigma}_m^2(x)\mathbb{1}_{\widehat{\sigma}_m^2(x)\leq\sqrt{L}} + \sqrt{L}\mathbb{1}_{\widehat{\sigma}_m^2(x)>\sqrt{L}} \tag{2.17}$$

where $L > 0$ is the same parameter used to define the approximation subspace $\mathcal{S}_{K_N,M}$. Thus, the risk of the ridge estimator $\widehat{\sigma}^2_{m,L}$ is upper-bounded as follows:

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L}-\sigma^2\right\|^2_{n,1}\right] &\leq \mathbb{E}\left[\left\|(\widehat{\sigma}^2_{m,L}-\sigma^2)\mathbb{1}_{[-\log(n),\log(n)]}\right\|^2_{n,1}\right]+\mathbb{E}\left[\left\|(\widehat{\sigma}^2_{m,L}-\sigma^2)\mathbb{1}_{[-\log(n),\log(n)]^c}\right\|^2_{n,1}\right]\\ &\leq \mathbb{E}\left[\left\|(\widehat{\sigma}^2_{m,L}-\sigma^2)\mathbb{1}_{[-\log(n),\log(n)]}\right\|^2_{n,1}\right]+4L\sup_{t\in[0,1]}\mathbb{P}(|X_t|>\log(n)).\end{aligned}$$

The first term on the *r.h.s.* is equivalent to the risk of a ridge estimator of σ^2 on the compact interval $[-\log(n),\log(n)]$. The second term on the *r.h.s.* is upper-bounded using Lemma 2.8.4. We derive below, an upper-bound of the risk of estimation of $\widehat{\sigma}^2_m$.

Theorem 2.3.4. *Suppose that $L=\log^2(n)$. Under Assumption 2.2.1 and for n large enough, it holds,*

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L}-\sigma^2\right\|^2_{n,1}\right]\leq \inf_{h\in\mathcal{S}_{m,L}}\|h-\sigma^2\|^2_n+C\sqrt{\frac{m^q\log^2(n)}{n}}$$

where $C>0$ is a constant, $q=1$ for the collection [**B**]*, and $q=2$ for the collection* [**F**].

We first remark that the upper-bound of the risk of the truncated estimator of σ^2 differs with respect to each of the chosen bases. This contrast comes from the fact that the Fourier basis $\{f_\ell,\ \ell=0,\dots,m-1\}$ and the spline basis $\{B_{\ell-M},\ \ell=0,\dots,m-1\}$ satisfy

$$\sum_{\ell=0}^{m-1}f_\ell(x)\leq C_f m,\ \text{and}\ \sum_{\ell=0}^{m-1}B_{\ell-M}(x)=1.$$

Secondly, the estimation error is not as fine as the one established in Theorem 2.3.2 where σ^2 is estimated on a compact interval. In fact, on the real line $\mathbb{R}$, the pseudo-norm $\|.\|_X$ can no longer be equivalent to the $\mathbb{L}^2-$norm since the transition density is not bounded from below on $\mathbb{R}$. Consequently, we cannot take advantage of the exact method used to establish the risk bound obtained in Theorem 2.3.2 which uses the equivalence relation between the pseudo-norms $\|.\|_{n,1}$ and $\|.\|_X$ on one side, and $\|.\|_X$ and the $\mathbb{L}^2-$norm $\|.\|$ on the other side. Moreover, we can also notice that the term of order $1/n^2$ does not appear since it is dominated by the estimation error.

We obtain below rates of convergence of the ridge estimator of σ^2 for each of the collections [**B**] and [**F**].

Corollary 2.3.5. *Suppose that $\sigma^2\in\Sigma_I(\beta,R)$ with $\beta\geq 1$*

For [B]. *Assume that $K\propto n^{1/(4\beta+1)}$. Under Assumptions 2.2.1 and for n large enough, there exists a constant $C>0$ depending on β and σ_1 such that*

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L}-\sigma^2\right\|^2_{n,1}\right]\leq C\log^{2\beta}(n)n^{-2\beta/(4\beta+1)}.$$

For [F]. *Assume that $m\propto n^{1/2(2\beta+1)}$. Under Assumptions 2.2.1 and for n large enough, it holds,*

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L}-\sigma^2\right\|^2_{n,1}\right]\leq C\log(n)n^{-\beta/(2\beta+1)}$$

where the constant $C>0$ depends on β and σ_1.

As we can remark, the obtained rates are slower than the ones established in Section 2.3.1 where σ^2 is estimated on a compact interval. This result is the immediate consequence of the result of Theorem 2.3.4.

2.4 Estimation of the diffusion coefficient from repeated diffusion paths

We now focus on the estimation of the (square) of the diffusion coefficient from i.i.d. discrete observations of the diffusion process ($N\to\infty$).

2.4.1 Non-adaptive estimation of the diffusion coefficient on a compact interval

We study the rate of convergence of the ridge estimators $\widehat{\sigma}^2_m$ of $\sigma^2_{|I}$ from $D_{N,n}$ when I is a compact interval. The next theorem gives an upper-bound of the risk of our estimators $\widehat{\sigma}^2_m,\ m \in \mathcal{M}$.

Theorem 2.4.1. *Suppose that $L = \log(Nn)$ and $\mathcal{M} = \left\{1, \ldots, \sqrt{\min(n,N)}/\log(Nn)\right\}$. Under Assumption 2.2.1 and for $n, N \to \infty$, for each $m \in \mathcal{M}$, there exist constants $C > 0$ and $C' > 0$ depending on σ_1 such that,*

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_m - \sigma^2_{|I}\right\|^2_{n,N}\right] \leq 3 \inf_{h \in \mathcal{S}_{m,L}} \left\|h - \sigma^2_{|I}\right\|^2_n + C\left(\frac{m}{Nn} + m\log(Nn)\exp\left(-C\sqrt{\min(n,N)}\right) + \Delta_n^2\right)$$
$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_m - \sigma^2_{|I}\right\|^2_{n}\right] \leq 34 \inf_{h \in \mathcal{S}_{m,L}} \left\|h - \sigma^2_{|I}\right\|^2_n + C'\left(\frac{m}{Nn} + m\log(Nn)\exp\left(-C\sqrt{\min(n,N)}\right) + \Delta_n^2\right).$$

Note that the result of Theorem 2.4.1 is independent of the choice of the basis that generates the approximation space $\mathcal{S}_m$. The first term on the right-hand side represents the approximation error of the initial space, the second term $\mathrm{O}\left(m/(Nn)\right)$ is the estimation error, and the last term characterizes the cost of the time discretization. The next result is derived from Theorem 2.4.1.

Corollary 2.4.2. *Suppose that $\sigma^2 \in \Sigma_I(\beta, R)$ with $\beta > 3/2$. Moreover, assume that $K_{\mathrm{opt}} \propto (Nn)^{1/(2\beta+1)}$ for $[\mathbf{B}]$ ($m_{\mathrm{opt}} = K_{\mathrm{opt}} + M$), and $m_{\mathrm{opt}} \propto (Nn)^{1/(2\beta+1)}$ for $[\mathbf{F}]$. Under Assumptions 2.2.1 and for $n, N \to \infty$, it holds,*

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m_{\mathrm{opt}}} - \sigma^2_{|I}\right\|^2_{n,N}\right] = \mathrm{O}\left((Nn)^{-2\beta/(2\beta+1)}\right)$$
$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m_{\mathrm{opt}}} - \sigma^2_{|I}\right\|^2_{n}\right] = \mathrm{O}\left((Nn)^{-2\beta/(2\beta+1)}\right).$$

The obtained result shows that the nonparametric estimators of $\sigma^2_{|I}$ based on repeated observations of the diffusion process are more efficient when $N, n \to \infty$. Note that the same rate is obtained if the risk of $\widehat{\sigma}^2_{m_{\mathrm{opt}}}$ is defined with the $\mathbb{L}^2-$norm $\|.\|$ equivalent to the empirical norm $\|.\|_n$.

The rate obtained in Corollary 2.4.2 is established for $\beta > 3/2$. If we consider for example the collection $[\mathbf{B}]$ and assume that $\beta \in [1, 3/2]$, then $K_{\mathrm{opt}} \propto (Nn)^{1/(2\beta+1)}$ belongs to $\mathcal{M}$ for $n \propto \sqrt{N}/\log^4(N)$ and we have

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m_{\mathrm{opt}}} - \sigma^2_{|I}\right\|^2_{n,N}\right] \leq C(Nn)^{-2\beta/(2\beta+1)}.$$

Under the condition $n \propto \sqrt{N}/\log^4(N)$ imposed on the length of diffusion paths, the obtained rate is of order $n^{-3\beta/(2\beta+1)}$ (up to a log-factor) which is equivalent to $N^{-3\beta/2(2\beta+1)}$ (up to a log-factor).

2.4.2 Non-adaptive estimation of the diffusion coefficient on the real line

Consider a ridge estimator of σ^2 on $\mathbb{R}$ built from N independent copies of the diffusion process X observed in discrete times, where both N and n tend to infinity. For each $m \in \mathcal{M}$, we still denote by $\widehat{\sigma}^2_m$ the ridge estimators of σ^2 and $\widehat{\sigma}^2_{m,L}$ the truncated estimators of σ^2 given in Equation (2.17). We establish, through the following theorem, the first risk bound that highlights the main error terms.

Theorem 2.4.3. *Suppose that $L = \log^2(N)$, and $N, n \to \infty$. Under Assumptions 2.2.1 and for any dimension $m \in \mathcal{M}$, the following holds:*

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L} - \sigma^2\right\|^2_{n,N}\right] \leq 2 \inf_{h \in \mathcal{S}_{m,L}} \|h - \sigma^2\|^2_n + C\left(\sqrt{\frac{m^q \log^2(N)}{Nn}} + \Delta_n^2\right)$$

where $C > 0$ is a constant depending on the upper bound σ_1 of the diffusion coefficient. Moreover, $q = 1$ for the collection $[\mathbf{B}]$ and $q = 2$ for the collection $[\mathbf{H}]$.

If we consider the risk of $\widehat{\sigma}^2_{m,L}$ using the empirical norm $\|.\|_n$, then we obtain

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L}-\sigma^2\right\|_n^2\right] \leq 2\inf_{h\in\mathcal{S}_{m,L}}\|h-\sigma^2\|_n^2 + C\left(\sqrt{\frac{m^q\log^2(N)}{Nn}}+\frac{m^2\log^3(N)}{N}+\Delta_n^2\right) \tag{2.18}$$

The risk bound given in Equation (2.18) is a sum of four error terms. The first term is the approximation error linked to the choice of the basis, the second term is the estimation error given in Theorem 2.4.3, the third term $m^2\log^3(N)/N$ comes from the relation linking the empirical norm $\|.\|_n$ to the pseudo-norm $\|.\|_{n,N}$ (see Lemma 2.8.7), and the last term is the cost of the time-discretization.

We derive, in the next result, rates of convergence of the risk bound of the truncated ridge estimators $\widehat{\sigma}^2_{m,L}$ based on the collections [**B**] and [**H**] respectively.

Corollary 2.4.4. *Suppose that $\sigma^2\in\Sigma_I(\beta,R)$ with $\beta\geq 1$, $I=[-\log(N),\log(N)]$, and $K\propto(Nn)^{1/(4\beta+1)}$ for* [**B**]*, and $\sigma^2\in W^s_{f_n}(\mathbb{R},R)$ with $s\geq 1$ and $m\propto(Nn)^{1/2(2s+1)}$ for* [**H**]*. Under Assumption 2.2.1 and for $N,n\to\infty$, the following holds:*

$$\text{For [B]}\ \ \mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L}-\sigma^2\right\|^2_{n,N}\right]\leq C\left(\log^{2\beta}(N)(Nn)^{-2\beta/(4\beta+1)}+\frac{1}{n^2}\right),$$

$$\text{For [H]}\ \ \mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L}-\sigma^2\right\|^2_{n,N}\right]\leq C\left(\log^{3}(N)(Nn)^{-s/(2s+1)}+\frac{1}{n^2}\right).$$

where $C>0$ is a constant depending on β and σ_1 for [**B**]*, or s and σ_1 for* [**H**].

The obtained rates are slower compared to the rates established in Section 2.4.1 for the estimation of $\sigma^2_{|I}$ where the interval $I\subset\mathbb{R}$ is compact. In fact, the method used to establish the rates of Theorem 2.4.3 from which the rates of Corollary 2.4.4 are obtained, does not allow us to derive rates of order $(Nn)^{-\alpha/(2\alpha+1)}$ (up to a log-factor) with $\alpha\geq 1$ (*e.g.* $\alpha=\beta,s$). Finally, if we consider the risk defined with the empirical norm $\|.\|_n$, then from Equation (2.18) with $n\propto N$ and assuming that $m\propto N^{1/4(s+1)}$ for [**H**] or $K\propto N^{1/4(\beta+1)}$ for [**B**], we obtain

$$\begin{aligned}
[\mathbf{B}]: &\quad \mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L}-\sigma^2\right\|_n^2\right]\leq C\log^{2\beta}(N)(Nn)^{-\beta/2(\beta+1)},\\
[\mathbf{H}]: &\quad \mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L}-\sigma^2\right\|_n^2\right]\leq C\log^{3}(N)(Nn)^{-s/2(s+1)},
\end{aligned}$$

where $C>0$ is a constant depending on σ_1 and on the smoothness parameter. We can see that the obtained rates are slower compared to the results of Corollary 2.4.4 for $n\propto N$. The deterioration of the rates comes from the additional term of order $m^2\log^3(N)/N$ which is now regarded as the new estimation error since it dominates the other term in each case as $N\to\infty$.

2.4.3 Non-adaptive estimation of the diffusion coefficient on a compact interval depending on the sample size

This section combines the two first sections 2.4.1 and 2.4.2 focusing on the estimation of σ^2 on the compact interval $[-A_N,A_N]$ where (A_N) is a strictly positive sequence such that $A_N\to\infty$ as $N\to\infty$. Consequently, we obtain that the estimation interval tends to $\mathbb{R}$ as the sample size N tends to infinity.

Define from the observations and for each dimension $m\in\mathcal{M}$, the following matrices:

$$\widehat{\Psi}_m := \left(\frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\phi_\ell\left(X^j_{k\Delta}\right)\phi_{\ell'}\left(X^j_{k\Delta}\right)\right)_{0\leq\ell,\ell'\leq m-1}, \tag{2.19}$$

$$\Psi_m := \mathbb{E}\left(\widehat{\Psi}_m\right) = \left(\mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1}\phi_\ell\left(X_{k\Delta}\right)\phi_{\ell'}\left(X_{k\Delta}\right)\right]\right)_{0\leq\ell,\ell'\leq m-1}. \tag{2.20}$$

These two matrices play an essential role in the construction of a consistent projection estimator of σ^2 over any approximation subspace $\mathcal{S}_m$ spanned by the basis $(\phi_0, \cdots, \phi_{m-1})$. Furthermore, for all $h = \sum_{\ell=0}^{m-1} a_\ell \phi_\ell \in \mathcal{S}_m$, we have:

$$\|h\|_{n,N}^2 = {}^t\mathbf{a}\widehat{\Psi}_m\mathbf{a}, \quad \|h\|_n^2 = \mathbb{E}\left(\|h\|_{n,N}^2\right) = {}^t\mathbf{a}\Psi_m\mathbf{a},$$

where $\mathbf{a} = (a_0, \cdots, a_{m-1})$. The Gram matrix Ψ_m is invertible under the spline basis (see [28]) and the Hermite basis (see [20]). We define for any invertible matrix M, the operator norm $\left\|M^{-1}\right\|_{\text{op}}$ of M^{-1} given by $\left\|M^{-1}\right\|_{\text{op}} = 1/\inf\{\lambda_j\}$ where the λ_j are eigenvalues of M.

For each dimension $m \in \mathcal{M}$, the matrices $\widehat{\Psi}_m$ and $\mathbf{F}_m$ satisfy:

$$\widehat{\Psi}_m = {}^t\mathbf{F}_m\mathbf{F}_m.$$

Consider the ridge estimator $\widehat{\sigma}_m^2$ of $\sigma_{A_N}^2 = \sigma^2\mathbb{1}_{[-A_N,A_N]}$, with $m \in \mathcal{M}$ and $A_N \to \infty$ as $N \to \infty$. The estimator $\widehat{\sigma}_m^2$ can reach a faster rate of convergence if the Gram matrix Ψ_m given in Equation (2.20) satisfies the following condition,

$$\mathcal{L}(m)\left(\left\|\Psi_m^{-1}\right\|_{\text{op}} \vee 1\right) \leq C\frac{N}{\log^2(N)}, \quad \text{where } \mathcal{L}(m) := \sup_{x\in\mathbb{R}}\sum_{\ell=0}^{m-1}\phi_\ell^2(x) < \infty \tag{2.21}$$

where $C > 0$ is a constant. In fact, the optimal rate of convergence is achieved on a random event $\Omega_{n,N,m}$ in which the two empirical norms $\|.\|_{n,N}$ and $\|.\|_n$ are equivalent (see [18], [30]). Then, Condition (2.21) is used to upper-bound $\mathbb{P}\left(\Omega_{n,N,m}^c\right)$ by a negligible term with respect to the considered rate (see [18]). Note that in Equation (2.21), the square on $\log(N)$ is justified by the fact that the value of constant $C > 0$ is unknown, and that the spline basis is not othonormal (see [28], proof of Lemma 7.8). The assumption of Equation (2.21) is also made in [18] on the operator norm of Ψ_m^{-1} based on an orthonormal basis with the bound $\mathbf{c}N/\log(N)$ where the value of $\mathbf{c}$ is known, and chosen and such that the upper-bound of $\mathbb{P}(\Omega_{n,N,m}^c)$ is negligible with respect to the estimation error. In our framework, since the transition density is approximated by Gaussian densities, we derive the following result.

Lemma 2.4.5. *Suppose that $n \propto N$ and $N \to \infty$, and that the spline basis is constructed on the interval $[-A_N, A_N]$ with $A_N > 0$. Under Assumption 2.2.1 , for all $m \in \mathcal{M}$ and for all $w \in \mathbb{R}^m$ such that $\|w\|_{2,m} = 1$, there exists a constant $C > 0$ such that*

$$\text{For } [\mathbf{H}]: \quad w'\Psi_m w \geq \frac{C}{\log(N)}\exp\left(-\frac{3c_\sigma(4m+3)}{2\left(1-\log^{-1}(N)\right)}\right),$$
$$\text{For } [\mathbf{B}]: \quad w'\Psi_m w \geq \frac{CA_N}{m\log(N)}\exp\left(-c_\sigma A_N^2\right),$$

where the constant $c_\sigma > 1$ that comes from the approximation of the transition density, depends on the diffusion coefficient σ.

The result of Lemma 2.4.5 implies for the Hermite basis that

$$\left(\left\|\Psi_m^{-1}\right\|_{\text{op}} \vee 1\right) \leq \frac{\log(N)}{C}\exp\left(\frac{3c_\sigma(4m+3)}{2\left(1-\log^{-1}(N)\right)}\right)$$

where the upper-bound is an exponentially increasing sequence of N since the dimension $m \in \mathcal{M}$ has a polynomial growth with respect to N. Thus, Condition (2.21) cannot be satisfied for the Hermite basis in our framework. Considering the spline basis, one has $\mathcal{L}(m) = \mathcal{L}(K+M) \leq 1$ and there exists a constant $C > 0$ such that

$$\left\|\Psi_m^{-1}\right\|_{\text{op}} \leq C\frac{m\log(N)}{A_N}\exp\left(c_\sigma A_N^2\right). \tag{2.22}$$

For $K \propto \left(N^{2/(2\beta+1)}A_N\right)$, Condition (2.21) is satisfied if the estimation interval $[-A_N, A_N]$ is chosen such that $A_N = o\left(\sqrt{\log(N)}\right)$. In the next theorem, we prove that the spline-based ridge estimator of $\sigma_{A_N}^2$ reaches a faster rate of convergence compared to the result of Corollary 2.4.4 for the collection $[\mathbf{B}]$.

Theorem 2.4.6. *Suppose that $N \propto n$ and $N \to \infty$, and consider the ridge estimator $\widehat{\sigma}^2_{A_N,m}$ of $\sigma^2_{A_N}$ based on the spline basis. Furthermore, suppose that $L = \log(N)$, $A_N = o\left(\sqrt{\log(N)}\right)$ and $K \propto (Nn)^{1/(2\beta+1)} A_N$ $(m = K + M)$. Under Assumptions 2.2.1 and for $\sigma^2 \in \Sigma_I(\beta, R)$ with $I = [-A_N, A_N]$, the following holds:*

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{A_N,m} - \sigma^2_{A_N}\right\|^2_{n,N}\right] \leq C \log^{\beta}(N)(Nn)^{-2\beta/(2\beta+1)}$$

where $C > 0$ is a constant depending on β.

The above result shows that the risk of the ridge estimator of $\sigma^2_{A_N}$ on $[-A_N, A_N]$ reaches a rate of order $(Nn)^{-\beta/(2\beta+1)}$ (up tp a log-factor) thanks to Condition (2.21) which allows us to take advantage of the equivalence relation between the empirical norms $\|.\|_n$ and $\|.\|_{n,N}$ given in Equation (2.3) to derive a finer estimation error (see proof of Theorem 2.4.6). Note that the obtained result depends on an appropriate choice of the estimation interval $[-A_N, A_N]$ which tends to $\mathbb{R}$ as N tends to infinity. Therefore, any choice of A_N such that $A_N/\sqrt{\log(N)} \longrightarrow +\infty$ cannot lead to a consistent estimation error since Equation (2.21) is no longer satisfied for the upper-bounding of $\mathbb{P}\left(\Omega^c_{n,N,m}\right)$ by a term that tends to zero as $N \to \infty$. Thus, the assumption $A_N = o\left(\sqrt{\log(N)}\right)$ is a necessary and sufficient condition for the validation of Condition (2.21) which leads, together with Assumption 2.2.1, to the result of Theorem 2.4.6. Finally, under the assumptions of Theorem 2.4.6 and considering the risk of $\widehat{\sigma}^2_{A_N,m}$ based on the empirical norm $\|.\|_n$, we also obtain

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{A_N,m} - \sigma^2_{A_N}\right\|^2_{n}\right] = \mathrm{O}\left(\log^{\beta}(N)(Nn)^{-2\beta/(2\beta+1)}\right).$$

In fact, under Condition (2.21), the estimator $\widehat{\sigma}^2_{A_N,m}$ satisfies the results of Theorem 2.4.1 with $I = [-A_N, A_N]$ and $A_N = o\left(\sqrt{\log(N)}\right)$, which implies rates of the same order for the two empirical norms.

2.5 Adaptive estimation of the diffusion coefficient from repeated observations

In this section, we suppose that $n \propto N$ and we propose a adaptive ridge estimator of σ^2 by selecting an optimal dimension from the sample D_N. In fact, consider the estimator $\widehat{\sigma}^2_{\widehat{K},L}$ where $\widehat{K}$ satisfies:

$$\widehat{K} := \underset{K \in \mathcal{K}}{\arg\min}\left\{\gamma_{n,N}\left(\widehat{\sigma}^2_K\right) + \mathrm{pen}(K)\right\} \tag{2.23}$$

and the penalty function $\mathrm{pen} : K \mapsto \mathrm{pen}(K)$ is established using the chaining technique of [7]. We derive below the risk of the adaptive estimator of $\sigma^2_{|I}$ when the interval $I \subset \mathbb{R}$ is compact and the sample size $N \to \infty$.

Theorem 2.5.1. *Suppose that $N \propto n$, $L = \log(N)$, $N \to \infty$, and consider the collection* [**B**] *with*

$$K \in \mathcal{K} = \{2^q,\ q = 0, 1, \ldots, q_{\max}\} \subset \mathcal{M} = \left\{1, \ldots, \left\lfloor \sqrt{N}/\log(N) \right\rfloor\right\}.$$

Under Assumption 2.2.1 , there exists a constant $C > 0$ such that,

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{\widehat{K},L} - \sigma^2_{|I}\right\|^2_{n,N}\right] \leq 34 \inf_{K \in \mathcal{K}}\left\{\inf_{h \in \mathcal{S}_{K+M,L}} \|h - \sigma^2_{|I}\|^2_n + \mathrm{pen}(K)\right\} + \frac{C}{Nn}$$

where $\mathrm{pen}(K) = \kappa(K + M)\log(N)/Nn$ with $\kappa > 0$ a numerical constant.

We deduce from Corollary 2.4.2 and its assumptions that the adaptive estimator $\widehat{\sigma}^2_{\widehat{K},L}$ satisfies:

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{\widehat{K},L} - \sigma^2_{|I}\right\|^2_{n}\right] = \mathrm{O}\left((Nn)^{-2\beta/(2\beta+1)}\right).$$

This result is justified since the penalty term is of the same order (up to a log-factor) than the estimation error established in Theorem 2.4.1.

Considering the adaptive estimator of σ^2 on the real line $I = \mathbb{R}$ when the sample size $N \to \infty$, we obtain the following result.

Theorem 2.5.2. *Suppose that $N \propto n$, $L = \log(N)$, $N \to \infty$, and consider the collection* [**B**] *with*

$$K \in \mathcal{K} = \{2^q,\ q = 0, 1, \dots, q_{\max}\} \subset \mathcal{M} = \left\{ \left\lfloor \sqrt{N}/\log(N) \right\rfloor \right\}.$$

Under Assumption 2.2.1 and for N large enough, the exists a constant $C > 0$ such that,

$$\mathbb{E}\left[\left\| \widehat{\sigma}^2_{\widehat{K},L} - \sigma^2 \right\|^2_{n,N} \right] \leq 3 \inf_{K \in \mathcal{K}} \left\{ \inf_{h \in \mathcal{S}_{K+M,L}} \left\| h - \sigma^2 \right\|^2_n + \mathrm{pen}(K) \right\} + \frac{C}{Nn}.$$

where $\mathrm{pen}(K) = \kappa' \frac{(K+M)\log(N)}{Nn}$ *with $\kappa' > 0$ a numerical constant.*

We have a penalty term of the same order than the one obtained in Theorem 2.5.1 where σ^2 is estimated on a compact interval. One can deduce that the adaptive estimator reaches a rate of the same order than the rate of the non-adaptive estimator given in Corollary 2.4.4 for the collection [**B**].

If we consider the adaptive estimator of the compactly supported diffusion coefficient built from a single diffusion path, we obtain below an upper-bound of its risk of estimation.

Theorem 2.5.3. *Suppose that $N = 1,\ n \to \infty,\ L = \sqrt{\log(n)}$ and consider the collection* [**B**] *with*

$$K \in \mathcal{K} = \{2^q,\ q = 0, \dots, q_{\max}\} \subset \mathcal{M} = \left\{1, \dots, \sqrt{n}/\log(n)\right\}.$$

Under Assumption 2.2.1, it holds

$$\mathbb{E}\left[\left\| \widehat{\sigma}^2_{\widehat{K},L} - \sigma^2_{|I} \right\|^2_{n,1} \right] \leq 3 \inf_{K \in \mathcal{K}} \left\{ \inf_{h \in \mathcal{S}_{K+M,L}} \left\| h - \sigma^2_{|I} \right\|^2_n + \mathrm{pen}(K) \right\} + \frac{C}{n}.$$

where $C > 0$ is a constant depending on τ_0, and $\mathrm{pen}(K) = \kappa \frac{(K+M)\log(n)}{n}$ *with $\kappa > 0$ a numerical constant.*

We deduce from Theorem 2.5.3 that if we assume that $\sigma^2 \in \Sigma_I(\beta, R)$, then the adaptive estimator $\widehat{\sigma}^2_{\widehat{K},L}$ reaches a rate of order $n^{-\beta/(2\beta+1)}$ (up to a log-factor). The result of this theorem is almost a deduction of the result of Theorem 2.5.1, the slight difference being the use, in the proofs, of the local time of the process and the equivalence relation between the pseudo-norm $\|.\|_{n,1}$ with the pseudo-norm $\|.\|_X$ instead of the empirical norm $\|.\|_n$ considered in the proof of Theorem 2.5.1.

2.6 Numerical study

This section is devoted to the numerical study on a simulation scheme. Section 2.6.1 focuses on the presentation of the chosen diffusion models. In Section 2.6.2, we describe the scheme for the implementation of the ridge estimators. We mainly focus on the **B**-spline basis for the numerical study, and in Section 2.6.3, we add a numerical study on the performance of the Hermite-based ridge estimator of σ^2 on $\mathbb{R}$. Finally, we compare the efficiency of our estimator built on the real line $\mathbb{R}$ from a single path with that of the Nadaraya-Watson estimator proposed in [42].

2.6.1 Models and simulations

Recall that the time horizon is $T = 1$ and $X_0 = 0$. Consider the following diffusion models:

Model 1 Ornstein-Uhlenbeck: $b(x) = 1 - x,\ \ \sigma(x) = 1$

Model 2: $b(x) = 1 - x,\ \ \sigma(x) = 1 - x^2$

Model 3: $b(x) = 1 - x,\ \ \sigma(x) = \frac{1}{3+\sin(2\pi x)} + \cos^2\left(\frac{\pi}{2}x\right)$

Model 1 is the commonly used Ornstein−Uhlenbeck model, known to be a simple diffusion model satisfying Assumption 2.2.1. Model 2 does not satisfy Assumption 2.2.1. Model 3 satisfies Assumption 2.2.1 with a multimodal diffusion coefficient.

The size N of the sample D_N takes values in the set $\{1, 10, 100, 1000\}$ where the length n of paths varies in the set $\{100, 250, 500, 1000\}$. As we work with the spline basis, the dimension $m = K + M$ of the approximation space is chosen such that $M = 3$ and K takes values in $\mathcal{K} = \{2^p,\ p = 0, \cdots, 5\}$ so that the subspaces are nested inside each other. We are using `R` for the simulation of diffusion paths via the function `sde.sim` of `sde` package, (see [59] for more details on the simulation of SDEs).

2.6.2 Implementation of the ridge estimators

In this section, we assess the quality of estimation of the adaptive estimator $\widehat{\sigma}^2_{\widehat{m}}$ in each of the 3 models through the computation of its risk of estimation. We compare the performance of the adaptive estimator with that of the oracle estimator $\widehat{\sigma}^2_{m^*}$ where m^* is given by:

$$m^* := \underset{m\in\mathcal{M}}{\arg\min} \left\|\widehat{\sigma}^2_m - \sigma^2\right\|^2_{n,N}. \tag{2.24}$$

For the spline basis, we have $m^* = K^* + M$ with $M = 3$. Finally, we complete the numerical study with a representation of a set of 10 estimators of σ^2 for each of the 3 models.

We evaluate the MISE of the spline-based adaptive estimators $\widehat{\sigma}^2_{\widehat{K}}$ by repeating 100 times the following steps:

1. Simulate samples $D_{N,n}$ and $D_{N',n}$ with $N \in \{1, 10, 100, 1000\}$, $N' = 100$ and $n \in \{100, 250, 1000\}$.
2. For each $K \in \mathcal{K}$, and from $D_{N,n}$, compute estimators $\widehat{\sigma}^2_K$ given in Equations (2.10) and (2.11).
3. Select the optimal dimension $\widehat{K} \in \mathcal{K}$ using Equation (2.23) and compute K^* from Equation (2.24)
4. Using $D_{N',n}$, evaluate $\left\|\widehat{\sigma}^2_{\widehat{K}} - \sigma^2\right\|^2_{n,N'}$ and $\left\|\widehat{\sigma}^2_{K^*} - \sigma^2\right\|^2_{n,N'}$.

We deduce the risks of estimation considering the average values of $\left\|\widehat{\sigma}^2_{\widehat{m}} - \sigma^2\right\|^2_{n,N'}$ and $\left\|\widehat{\sigma}^2_{m^*} - \sigma^2\right\|^2_{n,N'}$ over the 100 repetitions. Note that we consider in this section, the estimation of σ^2 on the compact interval $I = [-1, 1]$ and on the real line $\mathbb{R}$. The unknown parameters κ and κ' in the penalty functions given in Theorem 2.5.1 and Theorem 2.5.2 respectively, are numerically calibrated (details are given in Appendix 2.9), and we choose $\kappa = 4$ and $\kappa' = 5$ as their respective values.

2.6.3 Numerical results

We present in this section the numerical results of the performance of the spline-based adaptive estimators of $\sigma^2_{|I}$ with $I \subseteq \mathbb{R}$ together with the performance of the oracle estimators. We consider the case $I = [-1, 1]$ for the compactly supported diffusion coefficient, and the case $I = \mathbb{R}$.

Tables 2.2 and 2.3 present the numerical results of estimation of $\sigma^2_{|I}$ from simulated data following the steps given in Section 2.6.2.

The results of Table 2.2 and Table 2.3 show that the adapted estimator $\widehat{\sigma}^2_{\widehat{K}}$ is consistent, since its MISE tends to zero as both the size N of the sample $D_{N,n}$ and the length n of paths are larger. Moreover, note that in most cases, the ridge estimators of the compactly supported diffusion coefficients perform better than those of the non-compactly supported diffusion functions. As expected, we observe that the oracle estimator has generally a better performance compared to the adaptive estimator. Nonetheless, we can remark that the performances are very close in several cases, highlighting the efficiency of the data-driven selection of the dimension.

An additional important remark is the significant influence of the length n of paths on the performance of $\widehat{\sigma}^2_{\widehat{K}}$ and $\widehat{\sigma}^2_{K^*,L}$ (by comparison of Table 2.2 with Table 2.3), which means that estimators built from higher frequency data are more efficient. A similar remark is made for theoretical results obtained in Sections 2.4.2 and 2.4.1.

Performance of the Hermite-based estimator of the diffusion coefficient We focus on the estimation of σ^2 on $\mathbb{R}$ and assess the performance of its Hermite-based estimator (see Section 2.4.2). We present in Table 2.4, the performance of the oracle estimator $\widehat{\sigma}^2_{m^*,L}$.

From the numerical results of Table 2.4, we observe that the Hermite-based estimator of σ^2 is consistent as the sample size N and the length n paths take larger values.

Models	Intervals	Estimators	$N = 10$	$N = 100$	$N = 1000$
Model 1	$[-1, 1]$	$\widehat{\sigma}^2_{\widehat{K},L}$	0.0102 (0.0083)	0.0009 (0.0009)	0.0002 (0.0001)
		$\widehat{\sigma}^2_{K^*,L}$	0.0094 (0.0065)	0.0009 (0.0009)	0.0002 (0.0001)
	$\mathbb{R}$	$\widehat{\sigma}^2_{\widehat{K},L}$	0.0096 (0.0062)	0.0009 (0.0008)	0.0003 (0.0002)
		$\widehat{\sigma}^2_{K^*,L}$	0.0093 (0.0057)	0.0009 (0.0008)	0.0003 (0.0002)
Model 2	$[-1, 1]$	$\widehat{\sigma}^2_{\widehat{K},L}$	0.0048 (0.0052)	0.0019 (0.0008)	0.0005 (0.0002)
		$\widehat{\sigma}^2_{K^*,L}$	0.0039 (0.0043)	0.0009 (0.0005)	0.0005 (0.0002)
	$\mathbb{R}$	$\widehat{\sigma}^2_{\widehat{K},L}$	0.0195 (0.0140)	0.0057 (0.0006)	0.0012 (0.0002)
		$\widehat{\sigma}^2_{K^*,L}$	0.0048 (0.0064)	0.0025 (0.0021)	0.0010 (0.0003)
Model 3	$[-1, 1]$	$\widehat{\sigma}^2_{\widehat{K},L}$	0.0521 (0.0191)	0.0176 (0.0070)	0.0073 (0.0021)
		$\widehat{\sigma}^2_{K^*,L}$	0.0260 (0.0081)	0.0073 (0.0030)	0.0048 (0.0009)
	$\mathbb{R}$	$\widehat{\sigma}^2_{\widehat{K},L}$	0.1132 (0.0595)	0.0319 (0.0031)	0.0179 (0.0054)
		$\widehat{\sigma}^2_{K^*,L}$	0.0351 (0.0169)	0.0319 (0.0031)	0.0116 (0.0051)

Table 2.2: Assessment of MISEs (mean and standard deviation between brackets) of both the adaptive estimator $\widehat{\sigma}^2_{\widehat{K},L}$ and the oracle estimator $\widehat{\sigma}^2_{K^*,L}$ from diffusion paths of size $n = 100$.

Models	Intervals	Estimators	$N = 10$	$N = 100$	$N = 1000$
Model 1	$[-1, 1]$	$\widehat{\sigma}^2_{\widehat{K},L}$	0.0047 (0.0037)	0.0003 (0.0002)	0.0001 (0.00003)
		$\widehat{\sigma}^2_{K^*,L}$	0.0042 (0.0030)	0.0003 (0.0002)	0.0001 (0.00003)
	$\mathbb{R}$	$\widehat{\sigma}^2_{\widehat{K},L}$	0.0053 (0.0037)	0.0003 (0.0002)	0.0001 (0.00004)
		$\widehat{\sigma}^2_{K^*,L}$	0.0050 (0.0031)	0.0003 (0.0002)	0.0001 (0.00004)
Model 2	$[-1, 1]$	$\widehat{\sigma}^2_{\widehat{K},L}$	0.0027 (0.0019)	0.0003 (0.0002)	0.0002 (0.00004)
		$\widehat{\sigma}^2_{K^*,L}$	0.0018 (0.0019)	0.0002 (0.0001)	0.0001 (0.00004)
	$\mathbb{R}$	$\widehat{\sigma}^2_{\widehat{K},L}$	0.0091 (0.0077)	0.0028 (0.0025)	0.0008 (0.0002)
		$\widehat{\sigma}^2_{K^*,L}$	0.0020 (0.0023)	0.0021 (0.0023)	0.0002 (0.00004)
Model 3	$[-1, 1]$	$\widehat{\sigma}^2_{\widehat{K},L}$	0.0306 (0.0150)	0.0058 (0.0012)	0.0010 (0.0003)
		$\widehat{\sigma}^2_{K^*,L}$	0.0216 (0.0067)	0.0023 (0.0020)	0.0010 (0.0003)
	$\mathbb{R}$	$\widehat{\sigma}^2_{\widehat{K},L}$	0.0560 (0.0313)	0.0275 (0.0049)	0.0069 (0.0049)
		$\widehat{\sigma}^2_{K^*,L}$	0.0261 (0.0127)	0.0096 (0.0051)	0.0065 (0.0041)

Table 2.3: Assessment of MISEs of both the adaptive estimator $\widehat{\sigma}^2_{\widehat{K},L}$ and the oracle estimator $\widehat{\sigma}^2_{K^*,L}$ from diffusion paths of size $n = 250$.

Models	Intervals	Estimators	$N = 10,\ n = 100$	$N = 100,\ n = 100$	$N = 100,\ n = 250$
Model 1	$\mathbb{R}$	$\widehat{\sigma}^2_{K^*,L}$	0.0082 (0.0059)	0.0015 (0.0008)	0.0006 (0.0004)
Model 2	$\mathbb{R}$	$\widehat{\sigma}^2_{K^*,L}$	0.0058 (0.0111)	0.0007 (0.0004)	0.0003 (0.0002)
Model 3	$\mathbb{R}$	$\widehat{\sigma}^2_{K^*,L}$	0.0188 (0.0151)	0.0077 (0.0037)	0.0040 (0.0036)

Table 2.4: Assessment of MISEs of the Hermite-based oracle estimator $\widehat{\sigma}^2_{K^*,L}$ of the square of the diffusion coefficient.

Estimation of the diffusion coefficient from one path Consider ridge estimators of $\sigma^2_{|I}$ with $I = [-1, 1]$. For the case of the adaptive estimators of $\sigma^2_{|I}$, the dimension $\widehat{K}$ is selected such that

$$\widehat{K} = \arg\min_{K\in\mathcal{K}} \gamma_n(\widehat{\sigma}^2_{\widehat{K}}) + \text{pen}(K) \tag{2.25}$$

where $\text{pen}(K) = \kappa(K + M)\log(n)/n$ with $\kappa > 0$. We choose the numerical constant $\kappa = 4$ and we derive the numerical performance of the adaptive estimator of $\sigma^2_{|I}$.

Models	Intervals	Estimators	$n = 100$	$n = 1000$
Model 1	$[-1, 1]$	$\widehat{\sigma}^2_{\widehat{K},L}$	0.1751 (0.1921)	0.0915 (0.1925)
		$\widehat{\sigma}^2_{K^*,L}$	0.1563 (0.1776)	0.0783 (0.1699)
Model 2	$[-1, 1]$	$\widehat{\sigma}^2_{\widehat{K},L}$	0.1721 (0.3483)	0.1365 (0.5905)
		$\widehat{\sigma}^2_{K^*,L}$	0.0987 (0.1644)	0.0552 (0.2409)
Model 3	$[-1, 1]$	$\widehat{\sigma}^2_{\widehat{K},L}$	0.2184 (0.2780)	0.2106 (0.5790)
		$\widehat{\sigma}^2_{K^*,L}$	0.1263 (0.1486)	0.0751 (0.1469)

Table 2.5: Evaluation of MISEs of adaptive estimators $\widehat{\sigma}^2_{\widehat{K},L}$ built from a single diffusion path ($N = 1$) for each of the three models.

Table 2.5 gives the numerical performances of both the adaptive estimator and the oracle estimator of $\sigma^2_{|I}$ on the compact interval $I = [-1, 1]$ and from a single diffusion path. From the obtained results, we see that the estimators are numerically consistent. However, we note that the convergence is slow (increasing n from 100 to 1000), which highlights the significant impact of the number N of paths on the efficiency of the ridge estimator.

Comparison of the efficiency of the ridge estimator of the diffusion coefficient with its Nadaraya-Watson estimator. Consider the adaptive estimator $\widehat{\sigma}^2_{\widehat{K}}$ of the square of the diffusion coefficient buit on the real line $\mathbb{R}$ from a single diffusion path ($N = 1$), where the dimension $\widehat{K}$ is selected using Equation (2.25). For the numerical assessment, we use the interval $I = [-10^6, 10^6]$ to approximate the real line $\mathbb{R}$, and then, use Equation (2.25) for the data-driven selection of the dimension.

We want to compare the efficiency of $\widehat{\sigma}^2_{\widehat{K}}$ with that of the Nadaraya-Watson estimator of σ^2 given from a diffusion path $\bar{X} = (X_{k/n})_{1\le k\le n}$ and for all $x \in \mathbb{R}$ by

$$S_n(x) = \frac{\sum_{k=1}^{n-1} K\left(\frac{X_{k/n}-x}{h_n}\right)[X_{(k+1)/n} - X_{k/n}]^2/n}{\sum_{k=1}^{n} K\left(\frac{X_{k/n}-x}{h_n}\right)}$$

where K is a positive kernel function, and h_n is the bandwidth. Thus, the estimator $S_n(x)$ is consistent under the condition $nh_n^4 \to 0$ as n tends to infinity (see [42]). We use the function `ksdiff()` of the R-package `sde` to compute the Nadaraya−Watson estimator S_n.

Models	Ridge estimator	Nadaraya-Watson estimator
Model 1	0.0020 (0.0023)	0.9377 (0.0017)
Model 2	0.1323 (0.0794)	0.5086 (0.0885)
Model 3	0.4077 (0.1178)	1.3175 (0.3039)

Table 2.6: This table shows the loss errors of the ridge estimator $\widehat{\sigma}^2_{\widehat{K},L}$ on $\mathbb{R}$ and the Nadaraya-Watson estimator S_n of σ^2 built from a diffusion path ($N = 1$) of length $n = 1000$.

We remark from the results of Table 2.6 that our ridge estimator is more efficient. Note that for the kernel estimator S_n, the bandwidth is computed using the rule of thumb of Scott (see [75]). The bandwidth is proportional to $n^{-1/(d+4)}$ where n is the number of points, and d is the number of spatial dimensions.

2.6.4 Concluding remarks

The results of our numerical study show that our ridge estimators built both on a compact interval and on the real line are consistent as N and n take larger values, or as only n takes larger values when the estimators are built from a single path. These results are in accordance with the theoretical results established in the previous sections. Moreover, as expected, we obtained the consistency of the Hermite-based estimators of σ^2 on the real line $\mathbb{R}$. Nonetheless, we only focus on the Hermite-based oracle estimator since we did not establish a risk bound of the corresponding adaptive estimator. Finally, we remark that the ridge estimator of σ^2 built from a single path performs better than its Nadaraya-Watson kernel estimator proposed in [42] and implemented in the R-package `sde`.

2.7 Conclusion

In this article, we have proposed ridge-type estimators of the diffusion coefficient on a compact interval from a single diffusion path. We took advantage of the local time of the diffusion process to prove the consistency of non-adaptive estimators of σ^2 and derive a rate of convergence of the same order than the optimal rate established in [58]. We also propose an estimator of σ^2 on the real line from a single path. We proved its consistency using the method described in Section 2.3.2, and derive a rate of convergence order $n^{-\beta/(4\beta+1)}$ over a Hölder space for the collection [**B**]. Then, we extended the study to the estimation of σ^2 from repeated discrete observations of the diffusion process. We establish rates of convergence of the ridge estimators both on a compact interval and on $\mathbb{R}$. We complete the study proposing adaptive estimators of σ^2 on a compact interval for $N = 1$ and $N \to \infty$, and on the real line $\mathbb{R}$ for $N \to \infty$.

A perspective on the estimation of the diffusion coefficient could be the establishment of a minimax rate of convergence of the compactly supported (square of the) diffusion coefficient from repeated discrete observations of the diffusion process. The case of the non-compactly supported diffusion coefficient may be a lot more challenging, since the transition density of the diffusion process is no longer lower-bounded. This new fact can lead to different rates of convergence depending on the considered method (see Section 2.4).

2.8 Proofs

In this section, we prove our main results of Sections 2.3, 2.4 and 2.5. To simplify our notations, we set $\Delta_n = \Delta(= 1/n)$ and constants are generally denoted by $C > 0$ or $c > 0$ whose values can change from a line to another. Moreover, we use the notation C_α in case we need to specify the dependency of the constant C on a parameter α.

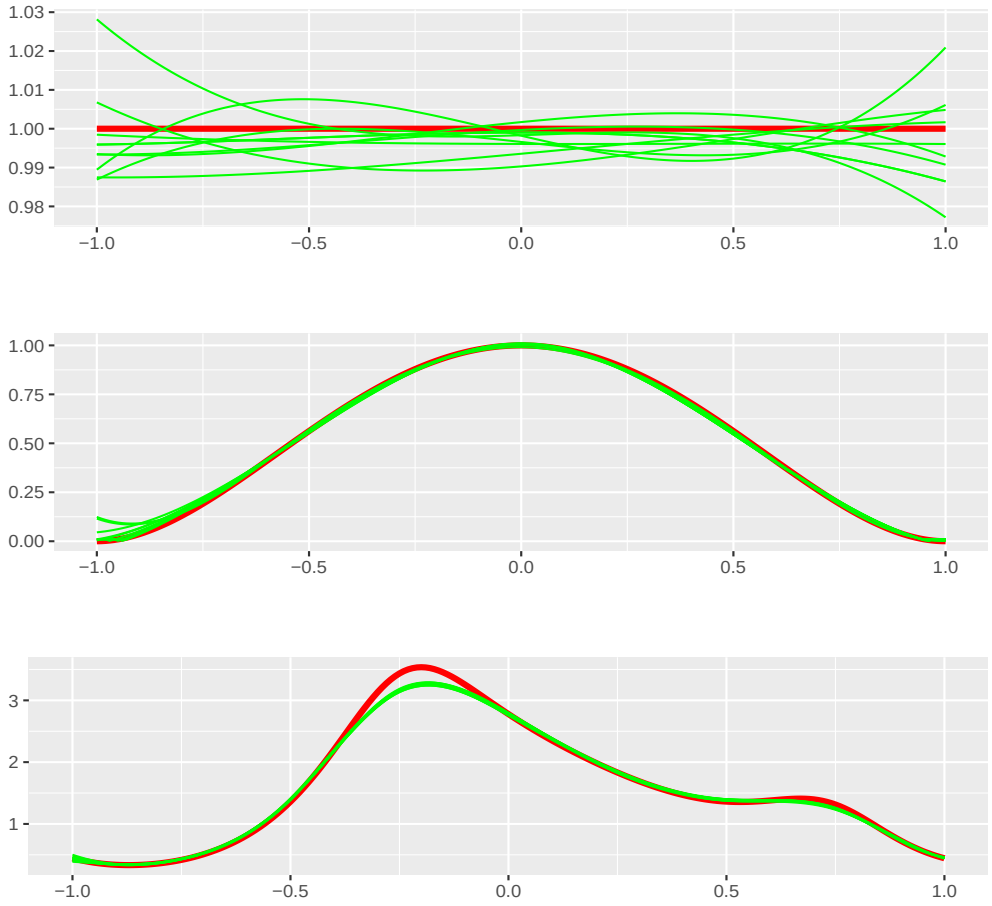

Figure 2.1: Bundles of 10 estimators $\widehat{\sigma}^2_{\widehat{K},L}$ (in green) of the true diffusion coefficient $\sigma^2_{|I}$ restricted on the compact interval $I = [-1, 1]$ (in red) of each of Models $1, 2, 3$ (from top to bottom) using samples of size $N = 1000$ with diffusion paths of length $n = 500$.

2.8.1 Technical results

Recall first some useful results on the local time and estimates of the transition density of diffusion processes.

Lemma 2.8.1. *For all integer $q \geq 1$, there exists $C^* > 0$ depending on q such that for all $0 \leq s < t \leq 1$,*

$$\mathbb{E}\left[|X_t - X_s|^{2q}\right] \leq C^*(t-s)^q.$$

The proof of Lemma 2.8.1 is provided in [28].

Proposition 2.8.2. *Under Assumptions 2.2.1, there exist constants $c_\sigma > 1$, $C > 1$ such that for all $t \in (0,1]$, $x \in \mathbb{R}$,*

$$\frac{1}{C\sqrt{t}} \exp\left(-c_\sigma \frac{x^2}{t}\right) \leq p_X(t,x) \leq \frac{C}{\sqrt{t}} \exp\left(-\frac{x^2}{c_\sigma t}\right).$$

The proof of Proposition 2.8.2 is provided in [52], *Proposition 1.2*.

Proposition 2.8.3. *Let h be a L_0-Lipschitz function. Then there exists $\tilde{h} \in \mathcal{S}_{K_N,M}$, such that*

$$|\tilde{h}(x) - h(x)| \leq C\frac{\log(N)}{K_N}, \quad \forall x \in (-\log(N), \log(N)),$$

where $C > 0$ depends on L_0, and M.

The proof of Proposition 2.8.3 is provided in [28]. The finite-dimensional vector space $\mathcal{S}_{K_N,M} = \mathcal{S}_{K_N+M}$ is introduced in Section 2.2.

Lemma 2.8.4. *Under Assumption 2.2.1, there exist $C_1, C_2 > 0$ such that for all $A > 0$,*

$$\sup_{t\in[0,1]} \mathbb{P}\left(|X_t| \geq A\right) \leq \frac{C_1}{A}\exp(-C_2A^2).$$

The proof of Lemma 2.8.4 is provided in [28], *Lemma 7.3*.

Lemma 2.8.5. *Under Assumption 2.2.1, the following holds:*

$$\forall\, x \in \mathbb{R}, \quad \mathcal{L}^x = \mathcal{L}^{x-} \quad a.s.$$

where $\mathcal{L}^{x-} = \lim_{\varepsilon\to 0}\mathcal{L}^{x-\varepsilon}$.

The result of Lemma 2.8.5 justifies the definition of the local time $\mathcal{L}^x$, for $x \in \mathbb{R}$, given in Equation (2.15).

Proof. From [81], *Theorem 1.7*, we have

$$\forall\, x \in \mathbb{R}, \quad \mathcal{L}^x - \mathcal{L}^{x-} = 2\int_0^1 \mathbb{1}_{X_s=x} dX_s = 2\int_0^1 \mathbb{1}_{X_s=x} b(X_s)ds + 2\int_0^1 \mathbb{1}_{X_s=x}\sigma(X_s)dW_s.$$

For each $s \in [0,1]$, let F_s be the density cumulative function of the random variable X_s. For all $x \in \mathbb{R}$ and for all $\varepsilon > 0$, , we have

$$\begin{aligned}\mathbb{P}(X_s = x) &= \lim_{\varepsilon\to 0}\mathbb{P}(X_s \leq x+\varepsilon) - \lim_{\varepsilon\to 0}\mathbb{P}(X_s \leq x-\varepsilon) = \lim_{\varepsilon\to 0} F_s(x+\varepsilon) - \lim_{\varepsilon\to 0} F_s(x-\varepsilon)\\ &= F_s(x) - F_s(x^-)\\ &= 0\end{aligned}$$

Thus, for all $x \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\left[|\mathcal{L}^x - \mathcal{L}^{x-}|\right] &\leq 2\int_0^1 |b(x)|\mathbb{P}(X_s = x)ds + 2\mathbb{E}\left[\left|\int_0^1 \mathbb{1}_{X_s=x}\sigma(X_s)dW_s\right|\right]\\ &= 2\mathbb{E}\left[\left|\int_0^1 \mathbb{1}_{X_s=x}\sigma(X_s)dW_s\right|\right].\end{aligned}$$

Using the Cauchy Schwartz inequality, we conclude that

$$\mathbb{E}\left[|\mathcal{L}^x - \mathcal{L}^{x-}|\right] \leq 2\sqrt{\mathbb{E}\left(\int_0^1 \mathbb{1}_{X_s=x}\sigma^2(X_s)ds\right)} = 2\sigma(x)\int_0^1 \mathbb{P}(X_s = x)ds = 0.$$

Using the Markov inequality, we have

$$\forall\, \varepsilon > 0,\ \ \mathbb{P}(|\mathcal{L}^x - \mathcal{L}^{x-}| > \varepsilon) \leq \frac{1}{\varepsilon}\mathbb{E}\left[|\mathcal{L}^x - \mathcal{L}^{x-}|\right] = 0.$$

We finally conclude that for all $x \in \mathbb{R}$,

$$\mathbb{P}(\mathcal{L}^x \neq \mathcal{L}^{x-}) = \mathbb{P}(|\mathcal{L}^x - \mathcal{L}^{x-}| > 0) = 0.$$

□

2.8.2 Proofs of Section 2.3

Proof of Lemma 2.3.1

Proof. The proof is divided into two parts for each of the two results to be proven.

First result. Since the function h is continuous on $\mathbb{R}$, let H be a primitive of h on $\mathbb{R}$. We deduce that for all $s \in [0,1]$,

$$h(X_s) = \lim_{\varepsilon\to 0}\frac{H(X_s+\varepsilon) - H(X_s-\varepsilon)}{2\varepsilon} = \lim_{\varepsilon\to 0}\frac{1}{2\varepsilon}\int_{X_s-\varepsilon}^{X_s+\varepsilon} h(x)dx = \lim_{\varepsilon\to 0}\frac{1}{2\varepsilon}\int_{-\infty}^{+\infty} h(x)\mathbb{1}_{(x-\varepsilon,x+\varepsilon)}(X_s)dx.$$

Finally, since h is integrable on $\mathbb{R}$ and using the theorem of dominated convergence, we obtain

$$\int_0^1 h(X_s)ds = \int_{-\infty}^{+\infty} h(x)\lim_{\varepsilon\to 0}\frac{1}{2\varepsilon}\int_0^1 \mathbb{1}_{(x-\varepsilon,x+\varepsilon)}(X_s)dsdx = \int_{-\infty}^{+\infty} h(x)\mathcal{L}^x dx.$$

Second result. Fix $t \in (0,1]$ and consider $P_X : (t,x) \mapsto \int_{-\infty}^x p_X(t,y)dy$ the cumulative density function of the random variable X_t of the density function $x \mapsto p_X(t,x)$. We have:

$$\begin{aligned}\forall\, x \in \mathbb{R},\ \ \mathbb{E}(\mathcal{L}^x) =& \lim_{\varepsilon\to 0}\frac{1}{2\varepsilon}\int_0^1 \mathbb{E}\left[\mathbb{1}_{(x-\varepsilon,x+\varepsilon)}(X_s)\right]ds = \lim_{\varepsilon\to 0}\frac{1}{2\varepsilon}\int_0^1 \mathbb{P}\left(x-\varepsilon \leq X_s \leq x+\varepsilon\right)ds \\ =& \int_0^1 \lim_{\varepsilon\to 0}\frac{P_X(s,x+\varepsilon) - P_X(s,x-\varepsilon)}{2\varepsilon}ds \\ =& \int_0^1 p_X(s,x)ds.\end{aligned}$$

□

Proof of Theorem 2.3.2

Let $\Omega_{n,m}$ be the random event in which the two pseudo-norms $\|.\|_{n,1}$ and $\|.\|_X$ are equivalent. $\Omega_{n,m}$ is given as follows:

$$\Omega_{n,m} := \bigcap_{g\in\mathcal{S}_m\setminus\{0\}}\left\{\left|\frac{\|g\|_{n,1}^2}{\|g\|_X^2} - 1\right| \leq \frac{1}{2}\right\}.$$

The proof of Theorem 2.3.2 relies on the following lemma.

Lemma 2.8.6. *Let $\gamma > 1$ be a real number. Under Assumption 2.2.1, the following holds*

$$\mathbb{P}\left(\Omega_{n,m}^c\right) \leq C\frac{m^{2\gamma}}{n^{\gamma/2}},$$

where $C > 0$ is a constant depending on γ.

The parameter $\gamma > 1$ has to be chosen appropriately (i.e. such that $m^{2\gamma}/n^{\gamma/2} = o(1/n)$) so that we obtain a variance term of the risk of the estimator $\widehat{\sigma}^2_m$ of order $m\log(n)/n$ (see Theorem 2.3.2 and Corollary 2.3.3).

Proof of Theorem 2.3.2 . Recall that since $N = 1$, $\zeta^1_{k\Delta} = \zeta^{1,1}_{k\Delta} + \zeta^{1,2}_{k\Delta} + \zeta^{1,3}_{k\Delta}$ is the error term of the regression model (1.6), with:

$$\zeta^{1,1}_{k\Delta} = \frac{1}{\Delta}\left[\left(\int_{k\Delta}^{(k+1)\Delta} \sigma(X^1_s)dW^1_s\right)^2 - \int_{k\Delta}^{(k+1)\Delta} \sigma^2(X^1_s)ds\right], \tag{2.26}$$

$$\zeta^{1,2}_{k\Delta} = \frac{2}{\Delta}\int_{k\Delta}^{(k+1)\Delta} ((k+1)\Delta - s)\sigma'(X^1_s)\sigma^2(X^1_s)dW^1_s, \tag{2.27}$$

$$\zeta^{1,3}_{k\Delta} = 2b(X^1_{k\Delta})\int_{k\Delta}^{(k+1)\Delta} \sigma\left(X^1_s\right)dW^1_s. \tag{2.28}$$

Besides, $R^1_{k\Delta} = R^{1,1}_{k\Delta} + R^{1,2}_{k\Delta} + R^{1,3}_{k\Delta}$, with:

$$R^{1,1}_{k\Delta} = \frac{1}{\Delta}\left(\int_{k\Delta}^{(k+1)\Delta} b(X^1_s)ds\right)^2, \quad R^{1,2}_{k\Delta} = \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta} ((k+1)\Delta - s)\Phi(X^1_s)ds \tag{2.29}$$

$$R^{1,2}_{k\Delta} = \frac{2}{\Delta}\left(\int_{k\Delta}^{(k+1)\Delta} \left(b(X^1_s) - b(X^1_{k\Delta})\right)ds\right)\left(\int_{k\Delta}^{(k+1)\Delta} \sigma(X^1_s)dW^1_s\right) \tag{2.30}$$

where

$$\Phi := 2b\sigma'\sigma + \left[\sigma''\sigma + \left(\sigma'\right)^2\right]\sigma^2. \tag{2.31}$$

By definition of the projection estimator $\widehat{\sigma}^2_m$ for each $m \in \mathcal{M}$ (see Equation (2.10)), for all $h \in \mathcal{S}_{m,L}$, we have:

$$\gamma_{n,1}\left(\widehat{\sigma}^2_m\right) - \gamma_{n,1}(\sigma^2_{|I}) \leq \gamma_{n,1}(h) - \gamma_{n,1}(\sigma^2_{|I}). \tag{2.32}$$

Furthermore, for all $h \in \mathcal{S}_{m,L}$,

$$\gamma_{n,1}(h) - \gamma_{n,1}(\sigma^2_{|I}) = \left\|\sigma^2_{|I} - h\right\|^2_{n,1} + 2\nu_1(\sigma^2_{|I} - h) + 2\nu_2(\sigma^2_{|I} - h) + 2\nu_3(\sigma^2_{|I} - h) + 2\mu(\sigma^2_{|I} - h),$$

where,

$$\nu_i(h) = \frac{1}{n}\sum_{k=0}^{n-1} h(X^1_{k\Delta})\zeta^{1,i}_{k\Delta}, \quad i \in \{1,2,3\}, \quad \mu(h) = \frac{1}{n}\sum_{k=0}^{n-1} h(X^1_{k\Delta})R^1_{k\Delta}, \tag{2.33}$$

and $\zeta^{1,1}_{k\Delta}$, $\zeta^{1,2}_{k\Delta}$, $\zeta^{1,3}_{k\Delta}$ are given in Equations (2.26), (2.27), (2.28), and finally, $R^1_{k\Delta} = R^{1,1}_{k\Delta} + R^{1,2}_{k\Delta}$ given in Equations (2.29) and (2.30). Then, for all $m \in \mathcal{M}$, and for all $h \in \mathcal{S}_{m,L}$, we obtain from Equation (2.32) that

$$\left\|\widehat{\sigma}^2_m - \sigma^2_{|I}\right\|^2_{n,1} \leq \left\|h - \sigma^2_{|I}\right\|^2_{n,1} + 2\nu\left(\widehat{\sigma}^2_m - h\right) + 2\mu\left(\widehat{\sigma}^2_m - h\right), \quad \text{with } \nu = \nu_1 + \nu_2 + \nu_3. \tag{2.34}$$

Then, it comes,

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_m - \sigma^2_{|I}\right\|^2_{n,1}\mathbb{1}_{\Omega_{n,m}}\right] \leq \inf_{h\in\mathcal{S}_{m,L}}\left\|h - \sigma^2_{|I}\right\|^2_n + 2\mathbb{E}\left[\nu\left(\widehat{\sigma}^2_m - h\right)\mathbb{1}_{\Omega_{n,m}}\right] + 2\mathbb{E}\left[\mu\left(\widehat{\sigma}^2_m - h\right)\mathbb{1}_{\Omega_{n,m}}\right]. \tag{2.35}$$

Besides, for any $a, d > 0$, using the inequality $xy \leq \eta x^2 + y^2/\eta$ with $\eta = a, d$, we have,

$$\begin{cases} 2\nu\left(\widehat{\sigma}^2_m - h\right) \leq \frac{2}{a}\left\|\widehat{\sigma}^2_m - \sigma^2_{|I}\right\|^2_X + \frac{2}{a}\left\|h - \sigma^2_{|I}\right\|^2_X + a\sup_{h\in\mathcal{S}_m,\ \|h\|_X=1}\nu^2(h), \\ 2\mu\left(\widehat{\sigma}^2_m - h\right) \leq \frac{2}{d}\left\|\widehat{\sigma}^2_m - \sigma^2_{|I}\right\|^2_{n,1} + \frac{2}{d}\left\|h - \sigma^2_{|I}\right\|^2_{n,1} + \frac{d}{n}\sum_{k=1}^n (R^1_{k\Delta})^2. \end{cases} \tag{2.36}$$

Upper bound of $\frac{1}{n}\sum_{k=1}^{n}(R_{k\Delta}^1)^2$

We have:

$$\begin{aligned}
&\forall k \in [\![1,n]\!],\ R_{k\Delta}^1 = R_{k\Delta}^{1,1} + R_{k\Delta}^{1,2} + R_{k\Delta}^{1,3} \text{ with,}\\
&R_{k\Delta}^{1,1} = \frac{1}{\Delta}\left(\int_{k\Delta}^{(k+1)\Delta} b(X_s^1)ds\right)^2,\ \ R_{k\Delta}^{1,2} = \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}((k+1)\Delta - s)\Phi(X_s^1)ds\\
&R_{k\Delta}^{1,3} = \frac{2}{\Delta}\left(\int_{k\Delta}^{(k+1)\Delta}\left(b(X_s^1) - b(X_{k\Delta}^1)\right)ds\right)\left(\int_{k\Delta}^{(k+1)\Delta}\sigma(X_s^1)dW_s^1\right).
\end{aligned}$$

For all $k \in [\![1,n]\!]$, using the Cauchy-Schwarz inequality and Equation (2.2),

$$\mathbb{E}\left[\left|R_{k\Delta}^{1,1}\right|^2\right] \le \mathbb{E}\left[\left(\int_{k\Delta}^{(k+1)\Delta} b^2(X_{k\Delta}^1)ds\right)^2\right] \le \Delta\mathbb{E}\left[\int_{k\Delta}^{(k+1)\Delta} b^4(X_{k\Delta}^1)ds\right] \le C\Delta^2.$$

Consider now the term $R_{k\Delta}^{1,2}$. From Equation (2.31), we have $\Phi = 2b\sigma'\sigma + \left[\sigma''\sigma + (\sigma')^2\right]\sigma^2$ and according to Assumption 2.2.1, there exists a constant $C > 0$ depending on σ_1 and α such that

$$\left|\Phi(X_s^1)\right| \le C\left[(2+|X_s^1|)(1+|X_s^1|^\alpha) + (1+|X_s^1|^\alpha)^2\right].$$

Then, from Equation (2.2) and for all $s \in (0,1]$,

$$\mathbb{E}\left[\Phi^2(X_s^1)\right] \le C\sup_{s\in(0,1]}\mathbb{E}\left[(2+|X_s^1|)^2(1+|X_s^1|^\alpha)^2 + (1+|X_s^1|^\alpha)^4\right] < \infty$$

and

$$\mathbb{E}\left[\left|R_{k\Delta}^{1,2}\right|^2\right] \le \frac{1}{\Delta^2}\int_{k\Delta}^{(k+1)\Delta}((k+1)\Delta - s)^2 ds \int_{k\Delta}^{(k+1)\Delta}\mathbb{E}\left[\Phi^2\left(X_s^1\right)\right]ds \le C\Delta^2$$

Finally, under Assumption 2.2.1, from Equation (2.2) and using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathbb{E}\left[\left|R_{k\Delta}^{1,3}\right|^2\right] &\le \frac{4}{\Delta^2}\mathbb{E}\left[\Delta\int_{k\Delta}^{(k+1)\Delta} L_0^2\left|X_s^1 - X_{k\Delta}^1\right|^2 ds\left(\int_{k\Delta}^{(k+1)\Delta}\sigma(X_s^1)dW_s\right)^2\right]\\
&\le \frac{4}{\Delta}\sqrt{\mathbb{E}\left[L_0^4\Delta\int_{k\Delta}^{(k+1)\Delta}\left|X_s^1 - X_{k\Delta}^1\right|^4 ds\right]\mathbb{E}\left[\left(\int_{k\Delta}^{(k+1)\Delta}\sigma(X_s^1)dW_s\right)^4\right]}\\
&\le C\Delta^2.
\end{aligned}$$

As a result, there exists a constant $C > 0$ such that,

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}(R_{k\Delta}^1)^2\right] \le C\Delta^2. \tag{2.37}$$

We set $a = d = 8$ and considering the event $\Omega_{n,m}$ on which the empirical norms $\|.\|_X$ and $\|.\|_{n,1}$ are equivalent, we deduce from Equations (2.35), (2.36) and (2.37) that,

$$\mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_{n,1}^2 \mathbb{1}_{\Omega_{n,m}}\right] \le 3\inf_{h\in\mathcal{S}_m}\left\|h - \sigma_{|I}^2\right\|_n^2 + C\mathbb{E}\left(\sup_{h\in\mathcal{S}_m,\|h\|_X=1}\nu^2(h)\right) + C\Delta^2 \tag{2.38}$$

where $C > 0$ is a constant depending on σ_1.

Upper bound of $\mathbb{E}\left(\sup_{h\in\mathcal{S}_m,\ \|h\|_X=1}\nu^2(h)\right)$

For all $h=\sum_{\ell=0}^{m-1}a_\ell\phi_\ell\in\mathcal{S}_m$ such that $\|h\|_X^2=1$, we have $\|h\|^2\leq\frac{1}{\tau_0}$ (see Equation (2.16)) and the coordinate vector $\mathbf{a}=(a_{-M},\cdots,a_{K-1})$ satisfies:

- $\|\mathbf{a}\|_2^2\leq Cm\ \ (m=K+M)$ for the spline basis (see [30], Lemma 2.6)
- $\|\mathbf{a}\|_2^2\leq 1/\tau_0$ for an orthonormal basis since $\|h\|^2=\|\mathbf{a}\|_2^2$.

Furthermore, using the Cauchy-Schwarz inequality, we have:

$$\nu^2(h)=\left(\sum_{\ell=0}^{m-1}a_\ell\nu\left(\phi_\ell\right)\right)^2\leq\|\mathbf{a}\|_2^2\sum_{\ell=0}^{m-1}\nu^2\left(\phi_\ell\right). \tag{2.39}$$

Thus, since $\nu=\nu_1+\nu_2+\nu_3$, for all $\ell\in[\![-M,K-1]\!]$ and for all $i\in\{1,2,3\}$,

$$\mathbb{E}\left[\nu_i^2\left(\phi_\ell\right)\right]=\frac{1}{n^2}\mathbb{E}\left[\left(\sum_{k=0}^{n-1}\phi_\ell\left(X_{k\Delta}^1\right)\zeta_{k\Delta}^{1,i}\right)^2\right].$$

1. Case $i=1$

   Recall that $\zeta_{k\Delta}^{1,1}=\frac{1}{\Delta}\left[\left(\int_{k\Delta}^{(k+1)\Delta}\sigma(X_s^1)dW_s\right)^2-\int_{k\Delta}^{(k+1)\Delta}\sigma^2(X_s^1)ds\right]$ where $W=W^1$. We fix a initial time $s\in[0,1)$ and set $M_t^s=\int_s^t\sigma(X_u^1)dW_u,\ \forall t\geq s$. $(M_t^s)_{t\geq s}$ is a martingale and for all $t\in[s,1]$, we have:

   $$\langle M^s,M^s\rangle_t=\int_s^t\sigma^2\left(X_u^1\right)du.$$

   Then, $\zeta_{k\Delta}^{1,1}=\frac{1}{\Delta}\left(M_{(k+1)\Delta}^{k\Delta}\right)^2-\left\langle M^{k\Delta},M^{k\Delta}\right\rangle_{(k+1)\Delta}$ is also a $\mathcal{F}_{k\Delta}$-martingale, and, using the Burkholder-Davis-Gundy inequality, we obtain for all $k\in[\![0,n-1]\!]$,

   $$\mathbb{E}\left[\zeta_{k\Delta}^{1,1}|\mathcal{F}_{k\Delta}\right]=0,\quad\mathbb{E}\left[\left(\zeta_{k\Delta}^{1,1}\right)^2|\mathcal{F}_{k\Delta}\right]\leq\frac{C}{\Delta^2}\mathbb{E}\left[\left(\int_{k\Delta}^{(k+1)\Delta}\sigma^2(X_u^1)du\right)^2\right]\leq C\sigma_1^4. \tag{2.40}$$

   Then, using Equation (2.40) we have:

   $$\begin{aligned}\mathbb{E}\left[\nu_1^2\left(\phi_\ell\right)\right]&=\frac{1}{n^2}\mathbb{E}\left[\sum_{k=0}^{n-1}\phi_\ell^2\left(X_{k\Delta}^1\right)\left(\zeta_{k\Delta}^{1,1}\right)^2\right]=\frac{1}{n^2}\mathbb{E}\left[\sum_{k=0}^{n-1}\phi_\ell^2\left(X_{k\Delta}^1\right)\mathbb{E}\left[\left(\zeta_{k\Delta}^{1,1}\right)^2|\mathcal{F}_{k\Delta}\right]\right]\\&\leq\frac{C\sigma_1^4}{n^2}\mathbb{E}\left[\sum_{k=0}^{n-1}\phi_\ell^2\left(X_{k\Delta}^1\right)\right]\end{aligned}$$

   and,

   $$\sum_{\ell=0}^{m-1}\mathbb{E}\left[\nu_1^2\left(\phi_\ell\right)\right]\leq\frac{C\sigma_1^4}{n^2}\mathbb{E}\left[\sum_{k=0}^{n-1}\sum_{\ell=0}^{m-1}\phi_\ell^2\left(X_{k\Delta}^1\right)\right].$$

   One has:

   $$\begin{cases}\sum_{\ell=-M}^{K-1}B_\ell^2\left(X_{k\Delta}^1\right)\leq 1\ \text{ for the Spline basis }\ (m=K+M),\\ \sum_{\ell=0}^{m-1}\phi_\ell^2(X_{k\Delta}^1)\leq Cm\ \text{ for an orthonormal basis with }\ C=\max_{0\leq\ell\leq m-1}\|\phi_\ell\|_\infty^2.\end{cases} \tag{2.41}$$

   Thus, it comes that

- $\sum_{\ell=-M}^{K-1} \mathbb{E}\left[\nu_1^2\left(B_\ell\right)\right] \leq C/n$ for the Spline basis,
- $\sum_{\ell=0}^{m-1} \mathbb{E}\left[\nu_1^2\left(\phi_\ell\right)\right] \leq Cm/n$ for an orthonormal basis,

and,

$$\mathbb{E}\left(\sup_{h\in\mathcal{S}_m,\ \|h\|_X^2=1} \nu_1^2(h)\right) \leq C\frac{m}{n} \tag{2.42}$$

where $C>0$ is a constant depending on σ_1 and the basis.

2. Case $i=2$

We set $\eta(s)=k\Delta$ for $s\in[k\Delta,(k+1)\Delta)$, $k=0,\ldots,n-1$. Wa have

$$\zeta_{k\Delta}^{1,2}=\frac{2}{\Delta}\int_{k\Delta}^{(k+1)\Delta}\left((k+1)\Delta-s\right)\sigma'\left(X_s^1\right)\sigma^2\left(X_s^1\right)dW_s,$$

and

$$\begin{aligned}\mathbb{E}\left[\nu_2^2\left(\phi_\ell\right)\right] &= 4\mathbb{E}\left[\left(\sum_{k=0}^{n-1}\phi_\ell\left(X_{k\Delta}^1\right)\int_{k\Delta}^{(k+1)\Delta}(k+1)\Delta-s)\sigma'\left(X_s^1\right)\sigma^2\left(X_s^1\right)dW_s\right)^2\right]\\ &= 4\mathbb{E}\left[\left(\int_0^1\phi_\ell\left(X_{\eta(s)}^1\right)\left(\eta(s)+\Delta-s\right)\sigma'\left(X_s^1\right)\sigma^2\left(X_s^1\right)dW_s\right)^2\right]\\ &\leq C\sigma_1^4\Delta^2\mathbb{E}\left[\int_0^1\phi_\ell^2\left(X_{\eta(s)}^1\right)ds\right]\end{aligned}$$

where $C>0$ is a constant. We deduce for both the spline basis and any orthonormal basis that there exists a constant $C>0$ depending on σ_1 such that:

$$\mathbb{E}\left(\sup_{h\in\mathcal{S}_m,\ \|h\|_X^2=1}\nu_2^2(h)\right)\leq C\frac{m}{n^2}. \tag{2.43}$$

3. Case $i=3$

We have $\zeta_{k\Delta}^{1,3}=2b\left(X_{k\Delta}^1\right)\int_{k\Delta}^{(k+1)\Delta}\sigma\left(X_s^1\right)dW_s$ and,

$$\begin{aligned}\mathbb{E}\left[\nu_3^2\left(\phi_\ell\right)\right] &= \frac{4}{n^2}\mathbb{E}\left[\left(\int_0^1\phi_\ell\left(X_{\eta(s)}^1\right)b\left(X_{\eta(s)}^1\right)\sigma\left(X_s^1\right)dW_s\right)^2\right]\\ &\leq \frac{4\sigma_1^2}{n^2}\mathbb{E}\left[\int_0^1\phi_\ell^2\left(X_{\eta(s)}^1\right)b^2\left(X_{\eta(s)}^1\right)ds\right]\end{aligned}$$

Since for all $x\in\mathbb{R}$, $b^2(x)\leq C_0(1+x^2)$ and $\sup_{t\in[0,1]}\mathbb{E}\left(|X_t|^2\right)<\infty$, there exists a constant $C>0$ depending on σ_1 such that:

$$\mathbb{E}\left(\sup_{h\in\mathcal{S}_m,\ \|h\|_X^2=1}\nu_3^2(h)\right)\leq C\frac{m}{n^2}. \tag{2.44}$$

We finally obtain from Equations (2.42), (2.43) and (2.44) that there exists a constant $C>0$ depending on σ_1 such that:

$$\mathbb{E}\left(\sup_{h\in\mathcal{S}_m,\ \|h\|_X^2=1}\nu^2(h)\right)\leq C\frac{m}{n}. \tag{2.45}$$

We deduce from Equations (2.38) and (2.45) that there exists a constant $C>0$ depending on σ_1 such that,

$$\mathbb{E}\left[\|\widehat{\sigma}_m^2-\sigma_{|I}^2\|_{n,1}^2\mathbb{1}_{\Omega_{n,m}}\right]\leq 3\inf_{h\in\mathcal{S}_{m,L}}\|\sigma_{|I}^2-h\|_n^2+C\left(\frac{m}{n}+\Delta^2\right).$$

For n large enough, we have $\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\|_\infty^2 \leq 2mL$ since $\|\widehat{\sigma}_m^2\|_\infty \leq \sqrt{mL}$. Then, from Lemma 2.8.6 and for all $m \in \mathcal{M}$, there exists a constant $C > 0$ depending on σ_1 such that

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_{n,1}^2\right] &= \mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_{n,1}^2 \mathbb{1}_{\Omega_{n,m}}\right] + \mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_{n,1}^2 \mathbb{1}_{\Omega_{n,m}^c}\right] \\ &\leq \mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_{n,1}^2 \mathbb{1}_{\Omega_{n,m}}\right] + 2mL\mathbb{P}\left(\Omega_{n,m}^c\right) \\ &\leq 3\inf_{h\in\mathcal{S}_{m,L}} \|\sigma_{|I}^2 - h\|_n^2 + C\left(\frac{m}{n} + \frac{m^{2\gamma+1}L}{n^{\gamma/2}} + \Delta^2\right).\end{aligned}$$

Since the pseudo-norms $\|.\|_{n,1}$ and $\|.\|_X$ are equivalent on the event $\Omega_{n,m}$, then, using Lemma 2.8.6, there exists a constant $C > 0$ depending on σ_1 such that

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_X^2\right] =& \mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_X^2 \mathbb{1}_{\Omega_{n,m}}\right] + \mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_X^2 \mathbb{1}_{\Omega_{n,m}^c}\right] \\ \leq& 8\mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_{n,1}^2\right] + 10\inf_{h\in\mathcal{S}_m}\left\|\sigma_{|I}^2 - h\right\|_n^2 + 2mL\mathbb{P}\left(\Omega_{n,m}^c\right) \\ \leq& 34\inf_{h\in\mathcal{S}_{m,L}}\left\|h - \sigma_{|I}^2\right\|_n^2 + C\left(\frac{m}{n} + \frac{m^{2\gamma+1}L}{n^{\gamma/2}} + \Delta^2\right).\end{aligned}$$

Finally, since the estimator $\widehat{\sigma}_m^2$ is built from a diffusion path $\bar{X}^1$ independent of the diffusion process X, and from Equations (2.16) and (2.13), the pseudo-norm $\|.\|_X$ depending on the process X and the empirical norm $\|.\|_n$ are equivalent ($\forall\, h \in \mathbb{L}^2(I),\ \|h\|_n^2 \leq (\tau_1/\tau_0)\mathbb{E}\left[\|h\|_X^2\right]$), there exists a constant $C > 0$ depending on σ_1, τ_0 and τ_1 such that

$$\mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_n^2\right] \leq \frac{34\tau_1}{\tau_0}\inf_{h\in\mathcal{S}_{m,L}}\left\|h - \sigma_{|I}^2\right\|_n^2 + C\left(\frac{m}{n} + \frac{m^{2\gamma+1}L}{n^{\gamma/2}} + \Delta^2\right).$$

□

Proof of Lemma 2.8.6 . The proof of this Lemma mainly focuses on the spline basis and the Fourier basis based on functions cos and sin which are Lipschitz functions. Thus, for all $g = \sum_{\ell=0}^{m-1} a_\ell\phi_\ell \in \mathcal{S}_m$,

$$\left|\|g\|_{n,1}^2 - \|g\|_X^2\right| \leq \int_0^1 \left|g^2(X_{\eta(s)}) - g^2(X_s)\right| ds \leq 2\|g\|_\infty \int_0^1 \left|g(X_{\eta(s)}) - g(X_s)\right| ds. \tag{2.46}$$

From Equation (2.16), one has $\mathbb{E}\left[\|g\|_X^2\right] \geq \tau_0\|g\|^2$. Thus, if $\|g\|_X^2 = 1$, then $\|g\|^2 \leq 1/\tau_0$, and we deduce for all $g = \sum_{\ell=0}^{m-1} a_\ell\phi_\ell$ that there exists a constant $C > 0$ such that

- **Spline basis**: $\|g\|_\infty \leq \|a\|_2 \leq C\sqrt{m}$ (see [30])
- **Fourier basis**: $\|g\|_\infty \leq C\sqrt{m}$ since $\|g\| = \|a\|_2$ and $\sum_{\ell=0}^{m-1}\phi_\ell^2 = \mathrm{O}(m)$.

Moreover, each $g \in \mathcal{S}_m$ such that $\|g\|_X^2 = 1$ is a Lipschitz function with a Lipschitz coefficient $L_g = \mathrm{O}\left(m^{3/2}\right)$. For the spline basis, this result is obtained in [30], *proof of Lemma C.1* combined with *Lemma 2.6*. For the Fourier basis, for all $x, y \in I$ and using the Cauchy Schwarz inequality, we obtain

$$\begin{aligned}|g(x) - g(y)| &\leq \sum_{\ell=0}^{m-1} |a_\ell|.|\phi_\ell(x) - \phi_\ell(y)| \\ &\leq 2\pi m\sqrt{m}\|\mathbf{a}\|_2|x - y| \\ &\leq \frac{2\pi}{\tau_0} m\sqrt{m}|x - y|.\end{aligned}$$

Back to Equation (2.46), there exists a constant $C > 0$ such that

$$\left|\|g\|_{n,1}^2 - \|g\|_X^2\right| \leq Cm^2\int_0^1 |X_{\eta(s)} - X_s| ds \tag{2.47}$$

We have:

$$\Omega_{n,m}^c = \left\{ \omega \in \Omega,\ \exists g \in \mathcal{S}_m \setminus \{0\},\ \left| \frac{\|g\|_{n,1}^2}{\|g\|_X^2} - 1 \right| > \frac{1}{2} \right\},$$

and, using Equation (2.47), we obtain

$$\sup_{g \in \mathcal{S}_m \setminus \{0\}} \left| \frac{\|g\|_{n,1}^2}{\|g\|_X^2} - 1 \right| = \sup_{g \in \mathcal{S}_m,\ \|g\|_X^2 = 1} \left| \|g\|_{n,1}^2 - \|g\|_X \right| \leq Cm^2 \int_0^1 |X_{\eta(s)} - X_s| ds.$$

Finally, using the Markov inequality, the Hölder inequality, Equation (2.2), and Lemma 2.8.1, we conclude that

$$\begin{aligned} \mathbb{P}\left(\Omega_{n,m}^c\right) &\leq \mathbb{P}\left(Cm^2 \int_0^1 |X_{\eta(s)} - X_s| ds \geq \frac{1}{2} \right) \\ &\leq Cm^{2\gamma} \int_0^1 \mathbb{E}\left[|X_{\eta(s)} - X_s|^\gamma \right] ds \\ &\leq C \frac{m^{2\gamma}}{n^{\gamma/2}} \end{aligned}$$

with $\gamma \in (1, +\infty)$. □

Proof of Theorem 2.3.4

Proof. Since $L = \log^2(n)$, we have

$$\begin{aligned} \mathbb{E}\left[\left\| \widehat{\sigma}_{m,L}^2 - \sigma^2 \right\|_{n,1}^2 \right] &= \mathbb{E}\left[\left\| (\widehat{\sigma}_{m,L}^2 - \sigma^2) \mathbb{1}_{[-\log(n), \log(n)]} \right\|_{n,1}^2 \right] + \mathbb{E}\left[\left\| (\widehat{\sigma}_{m,L}^2 - \sigma^2) \mathbb{1}_{[-\log(n), \log(n)]^c} \right\|_{n,1}^2 \right] \\ &\leq \mathbb{E}\left[\left\| (\widehat{\sigma}_{m,L}^2 - \sigma^2) \mathbb{1}_{[-\log(n), \log(n)]} \right\|_{n,1}^2 \right] + 2\log^2(n) \sup_{t \in (0,1]} \mathbb{P}(|X_t| > \log(n)). \end{aligned}$$

From Equation (2.35) (Proof of Theorem 2.3.2), for all $h \in \mathcal{S}_{m,L}$,

$$\mathbb{E}\left[\left\| (\widehat{\sigma}_{m,L}^2 - \sigma^2) \mathbb{1}_{[-\log(n), \log(n)]} \right\|_{n,1}^2 \right] \leq \inf_{h \in \mathcal{S}_{m,L}} \left\| h - \sigma^2 \right\|_n^2 + 2 \sum_{i=1}^3 \mathbb{E}\left[\nu_i \left(\widehat{\sigma}_m^2 - h \right) \right] + 2\mathbb{E}\left[\mu \left(\widehat{\sigma}_m^2 - h \right) \right] \tag{2.48}$$

where $\nu_i,\ i = 1, 2, 3$ and μ are given in Equation (2.33). For all $i \in \{1, 2, 3\}$ and for all $h \in \mathcal{S}_{m,L}$, one has

$$\mathbb{E}\left[\nu_i \left(\widehat{\sigma}_{m,L}^2 - h \right) \right] \leq \sqrt{2m \log^2(n)} \sqrt{ \sum_{\ell=0}^{m-1} \mathbb{E}\left[\nu_i^2(\phi_\ell) \right] }. \tag{2.49}$$

1. Upper bound of $\sum_{\ell=0}^{m-1} \mathbb{E}\left[\nu_1^2(\phi_\ell) \right]$

   According to Equation (2.33), we have

   $$\forall \ell \in [\![0, m-1]\!],\ \nu_1(\phi_\ell) = \frac{1}{n} \sum_{k=0}^{n-1} \phi_\ell(X_{k\Delta}^1) \zeta_{k\Delta}^{1,1}$$

   where $\zeta_{k\Delta}^{1,1} = \frac{1}{\Delta} \left[\left(\int_{k\Delta}^{(k+1)\Delta} \sigma(X_s^1) dW_s \right)^2 - \int_{k\Delta}^{(k+1)\Delta} \sigma^2(X_s^1) ds \right]$ is a martingale satisfying

   $$\mathbb{E}\left[\zeta_{k\Delta}^{1,1} | \mathcal{F}_{k\Delta} \right] = 0 \ \text{ and } \ \mathbb{E}\left[\left(\zeta_{k\Delta}^{1,1} \right)^2 | \mathcal{F}_{k\Delta} \right] \leq \frac{1}{\Delta^2} \mathbb{E}\left[\left(\int_{k\Delta}^{(k+1)\Delta} \sigma^2(X_s^1) ds \right)^2 \right] \leq C\sigma_1^4$$

with $C > 0$ a constant, $W = W^1$ and $(\mathcal{F}_t)_{t\geq 0}$ the natural filtration of the martingale $(M_t)_{t\in[0,1]}$ given for all $t \in [0,1]$ by $M_t = \int_0^t \sigma(X_s^1)dW_s$. We derive that

$$\sum_{\ell=0}^{m-1} \mathbb{E}\left[\nu_1^2(\phi_\ell)\right] = \frac{1}{n^2}\sum_{\ell=0}^{m-1}\mathbb{E}\left[\left(\sum_{k=0}^{n-1}\phi_\ell(X_{k\Delta}^1)\zeta_{k\Delta}^{1,1}\right)^2\right] = \frac{1}{n^2}\mathbb{E}\left[\sum_{k=0}^{n-1}\sum_{\ell=0}^{m-1}\phi_\ell^2(X_{k\Delta}^1)\left(\zeta_{k\Delta}^{1,1}\right)^2\right]$$

since for all integers k, k' such that $k > k' \geq 0$, we have

$$\mathbb{E}\left[\phi_\ell(X_{k\Delta}^1)\zeta_{k\Delta}^{1,1}\phi_\ell(X_{k'\Delta}^1)\zeta_{k'\Delta}^{1,1}|\mathcal{F}_{k\Delta}\right] = \phi_\ell(X_{k\Delta}^1)\zeta_{k'\Delta}^{1,1}\phi_\ell(X_{k'\Delta}^1)\mathbb{E}\left[\zeta_{k\Delta}^{1,1}|\mathcal{F}_{k\Delta}\right] = 0.$$

For each $k \in [\![0, n-1]\!]$, we have

$$\begin{cases} \sum_{\ell=0}^{m-1}\phi_\ell(X_{k\Delta}^1) = \sum_{\ell=-M}^{K-1} B_\ell(X_{k\Delta}^1) = 1 \ \text{ for the spline basis} \\ \\ \sum_{\ell=0}^{m-1}\phi_\ell(X_{k\Delta}^1) \leq Cm \ \text{ For an orthonormal basis with } C = \max\limits_{0\leq\ell\leq m-1}\|\phi_\ell\|_\infty. \end{cases}$$

Finally, there exists a constant $C > 0$ such that

$$\sum_{\ell=0}^{m-1}\mathbb{E}\left[\nu_1^2(\phi_\ell)\right] \leq \begin{cases} \frac{C}{n} \ \text{ for the spline basis} \\ \\ C\frac{m}{n} \ \text{ for an orthonormal basis.} \end{cases}$$

2. Upper bound of $\sum_{\ell=0}^{m-1}\mathbb{E}\left[\nu_2^2(\phi_\ell)\right]$

   For all $k \in [\![0, n-1]\!]$ and for all $s \in [0,1]$, set $\eta(s) = k\Delta$ if $s \in [k\Delta, (k+1)\Delta)$. We have:

$$\begin{aligned}\sum_{\ell=0}^{m-1}\mathbb{E}\left[\nu_2^2(\phi_\ell)\right] &= 4\sum_{\ell=0}^{m-1}\mathbb{E}\left[\left(\sum_{k=0}^{n-1}\int_{k\Delta}^{(k+1)\Delta}\phi_\ell(X_{k\Delta}^1)((k+1)\Delta - s)\sigma'(X_s^1)\sigma^2(X_s^1)dW_s\right)^2\right] \\ &= 4\sum_{\ell=0}^{m-1}\mathbb{E}\left[\left(\int_0^1\phi_\ell(X_{\eta(s)}^1)(\eta(s)+\Delta-s)\sigma'(X_s^1)\sigma^2(X_s^1)dW_s\right)^2\right].\end{aligned}$$

   We conclude that

$$\sum_{\ell=0}^{m-1}\mathbb{E}\left[\nu_2^2(\phi_\ell)\right] \leq \begin{cases} \frac{C}{n^2} \ \text{ for the spline basis} \\ \\ C\frac{m}{n^2} \ \text{ for an orthonormal basis.} \end{cases}$$

   where the constant $C > 0$ depends on the diffusion coefficient and the upper bound of the basis functions.

3. Upper bound of $\sum_{\ell=0}^{m-1}\mathbb{E}\left[\nu_2^3(\phi_\ell)\right]$

   We have:

$$\begin{aligned}\sum_{\ell=0}^{m-1}\mathbb{E}\left[\nu_3^2(\phi_\ell)\right] &= \frac{4}{n^2}\sum_{\ell=0}^{m-1}\mathbb{E}\left[\left(\sum_{k=0}^{n-1}\int_{k\Delta}^{(k+1)\Delta}\phi_\ell(X_{k\Delta}^1)b(X_{k\Delta}^1)\sigma(X_s^1)dW_s\right)^2\right] \\ &= \frac{4}{n^2}\sum_{\ell=0}^{m-1}\mathbb{E}\left[\left(\int_0^1\phi_\ell(X_{\eta(s)}^1)b(X_{\eta(s)}^1)\sigma(X_s^1)dW_s\right)^2\right] \\ &\leq \frac{4}{n^2}\mathbb{E}\left[\int_0^1\sum_{\ell=0}^{m-1}\phi_\ell^2(X_{\eta(s)}^1)b^2(X_{\eta(s)}^1)\sigma^2(X_s^1)ds\right].\end{aligned}$$

Since for all $x \in \mathbb{R}$, $b(x) \leq C_0(1+x^2)$ and $\sup_{t\in[0,1]} \mathbb{E}\left(|X_t|^4\right) < \infty$, there exists a constant $C > 0$ depending on the diffusion coefficient such that

$$\sum_{\ell=0}^{m-1} \mathbb{E}\left[\nu_3^2(\phi_\ell)\right] \leq \begin{cases} \frac{C}{n^2} & \text{for the spline basis} \\ \\ C\frac{m}{n^2} & \text{for an orthonormal basis.} \end{cases}$$

We finally deduce that from Equations (2.48) and (2.49) that for all $h \in \mathcal{S}_{m,L}$,

$$\begin{cases} \mathbb{E}\left[\|(\widehat{\sigma}^2_{m,L} - \sigma^2)\mathbb{1}_{[-\log(n),\log(n)]}\|^2_{n,1}\right] \leq \inf_{h\in\mathcal{S}_m} \|h-\sigma^2\|_n^2 + C\sqrt{\frac{m\log^2(n)}{n}} + 2\mathbb{E}\left[\mu(\widehat{\sigma}^2_{m,L} - h)\right] & [\mathbf{B}] \\ \mathbb{E}\left[\|(\widehat{\sigma}^2_{m,L} - \sigma^2)\mathbb{1}_{[-\log(n),\log(n)]}\|^2_{n,1}\right] \leq \inf_{h\in\mathcal{S}_m} \|h-\sigma^2\|_n^2 + C\sqrt{\frac{m^2\log^2(n)}{n}} + 2\mathbb{E}\left[\mu(\widehat{\sigma}^2_{m,L} - h)\right] & [\mathbf{F}] \end{cases} \tag{2.50}$$

where $C > 0$ is a constant. It remains to obtain an upper bound of the term $\mu(\widehat{\sigma}^2_{m,L} - h)$. For all $a > 0$ and for all $h \in \mathcal{S}_{m,L}$,

$$\begin{aligned} 2\mu\left(\widehat{\sigma}^2_{m,L} - h\right) \leq & \frac{2}{a}\left\|\widehat{\sigma}^2_{m,L} - \sigma^2\right\|^2_{n,1} + \frac{2}{a}\left\|h-\sigma^2\right\|^2_{n,1} + \frac{a}{n}\sum_{k=0}^{n-1}\left(R^1_{k\Delta}\right)^2 \\ 2\mathbb{E}\left[\mu\left(\widehat{\sigma}^2_{m,L} - h\right)\right] \leq & \frac{2}{a}\mathbb{E}\left\|\widehat{\sigma}^2_{m,L} - \sigma^2\right\|^2_{n,1} + \frac{2}{a}\inf_{h\in\mathcal{S}_m}\|h-\sigma^2\|_n^2 + \frac{a}{n}\sum_{k=0}^{n-1}\mathbb{E}\left[\left(R^1_{k\Delta}\right)^2\right]. \end{aligned}$$

Using Equations (2.37), (2.50) and setting $a = 4$, we deduce that there exists constant $C > 0$ depending on σ_1 such that,

$$\begin{cases} \mathbb{E}\left[\|\widehat{\sigma}^2_{m,L} - \sigma^2\|^2_{n,1}\right] \leq \inf_{h\in\mathcal{S}_m} \|h-\sigma^2\|_n^2 + C\sqrt{\frac{m\log^2(n)}{n}} + 2\log^2(n)\sup_{t\in(0,1]}\mathbb{P}(|X_t| > A_n) & [\mathbf{B}] \\ \mathbb{E}\left[\|\widehat{\sigma}^2_{m,L} - \sigma^2\|^2_{n,1}\right] \leq \inf_{h\in\mathcal{S}_m} \|h-\sigma^2\|_n^2 + C\sqrt{\frac{m^2\log^2(n)}{n}} + 2\log^2(n)\sup_{t\in(0,1]}\mathbb{P}(|X_t| > A_n) & [\mathbf{F}]. \end{cases} \tag{2.51}$$

From Proposition 2.8.3, $\sup_{t\in(0,1]}\mathbb{P}(|X_t| > \log(n)) \leq \log^{-1}(n)\exp(-c\log^2(n))$ with $c > 0$ a constant. Then, we obtain from Equation (2.51) that

$$\begin{cases} \mathbb{E}\left[\|\widehat{\sigma}^2_{m,L} - \sigma^2\|^2_{n,1}\right] \leq \inf_{h\in\mathcal{S}_m} \|h-\sigma^2\|_n^2 + C\sqrt{\frac{m\log^2(n)}{n}} & [\mathbf{B}] \\ \\ \mathbb{E}\left[\|\widehat{\sigma}^2_{m,L} - \sigma^2\|^2_{n,1}\right] \leq \inf_{h\in\mathcal{S}_m} \|h-\sigma^2\|_n^2 + C\sqrt{\frac{m^2\log^2(n)}{n}} & [\mathbf{F}]. \end{cases} \tag{2.52}$$

□

2.8.3 Proof of Section 2.4

The following lemma allows us to obtain a risk bound of $\widehat{\sigma}^2_{m,L}$ defined with the empirical norm $\|.\|_n$ from the risk bound defined from the pseudo norm $\|.\|_{n,N}$.

Lemma 2.8.7. *Let $\widehat{\sigma}^2_{m,L}$ be the truncated projection estimator on $\mathbb{R}$ of σ^2 over the subspace $\mathcal{S}_{m,L}$. Suppose that $L = \log^2(N)$, $N > 1$. Under Assumption 2.2.1, there exists a constant $C > 0$ independent of m and N such that*

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L} - \sigma^2\right\|^2_{n,N}\right] - 2\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L} - \sigma^2\right\|^2_{n}\right] \leq C\frac{m^2\log^3(N)}{N}.$$

The proof of Lemma 2.8.7 is provided in [30], *Theorem 3.3*. The proof uses the independence of the copies $\bar{X}^1, \ldots, \bar{X}^N$ of the process X at discrete times, and the Bernstein inequality.

Proof of Theorem 2.4.1

For fixed n and N in $\mathbb{N}^*$, we set for all $m \in \mathcal{M}$,

$$\Omega_{n,N,m} := \bigcap_{h \in \mathcal{S}_m \setminus \{0\}} \left\{ \left| \frac{\|h\|_{n,N}^2}{\|h\|_n^2} - 1 \right| \leq \frac{1}{2} \right\}. \tag{2.53}$$

As we can see, the empirical norms $\|h\|_{n,N}$ and $\|h\|_n$ of any function $h \in \mathcal{S}_m \setminus \{0\}$ are equivalent on $\Omega_{n,N,m}$. More precisely, on the set $\Omega_{n,N,m}$, for all $h \in \mathcal{S}_m \setminus \{0\}$, we have : $\frac{1}{2}\|h\|_n^2 \leq \|h\|_{n,N}^2 \leq \frac{3}{2}\|h\|_n^2$. We have the following result:

Lemma 2.8.8. *Under Assumption 2.2.1, the following holds:*

- *If $n \geq N$ or $n \propto N$, then $m \in \mathcal{M} = \left\{1, \ldots, \left\lfloor \sqrt{N}/\log(Nn) \right\rfloor \right\}$ and,*

$$\mathbb{P}\left(\Omega_{n,N,m}^c\right) \leq 2\exp(-C\sqrt{N}).$$

- *If $n \leq N$, then $m \in \mathcal{M} = \{1, \ldots, \lfloor \sqrt{n}/\log(Nn) \rfloor\}$ and*

$$\mathbb{P}\left(\Omega_{n,N,m}^c\right) \leq 2\exp(-C\sqrt{n})$$

where $C > 0$ is a constant.

Proof of Lemma 2.8.8 . We have:

$$\Omega_{n,N,m}^c = \left\{ \omega \in \Omega,\ \exists h_0 \in \mathcal{S}_m,\ \left| \frac{\|h\|_{n,N}^2}{\|h\|_n^2} - 1 \right| > \frac{1}{2} \right\},$$

Denote by $\mathcal{H}_m = \{h \in \mathcal{S}_m,\ \|h\|_n = 1\}$ and $\mathcal{H}_m^\varepsilon$ the $\varepsilon-$net of $\mathcal{H}_m$ for any $\varepsilon > 0$. We have

$$\sup_{h \in \mathcal{H}_m} \left| \frac{\|h\|_{n,N}^2}{\|h\|_n^2} - 1 \right| = \sup_{h \in \mathcal{H}_m} \left| \|h\|_{n,N}^2 - 1 \right|.$$

Let $\varepsilon > 0$ and let $\mathcal{H}_m^\varepsilon$ be the $\varepsilon-$net of $\mathcal{H}_m$ *w.r.t.* the supremum norm $\|.\|_\infty$. Then, for each $h \in \mathcal{H}_m$, there exists $h_\varepsilon \in \mathcal{H}_m^\varepsilon$ such that $\|h - h_\varepsilon\|_\infty \leq \varepsilon$. Then

$$\left| \|h\|_{n,N}^2 - 1 \right| \leq \left| \|h\|_{n,N}^2 - \|h_\varepsilon\|_{n,N}^2 \right| + \left| \|h_\varepsilon\|_{n,N}^2 - 1 \right|$$

and,

$$\left| \|h\|_{n,N}^2 - \|h_\varepsilon\|_{n,N}^2 \right| \leq \frac{1}{Nn} \sum_{j=1}^{N} \sum_{k=0}^{n-1} \left| h(X_{k\Delta}^j) - h_\varepsilon(X_{k\Delta}^j) \right| (\|h\|_\infty + \|h_\varepsilon\|_\infty) \leq (\|h\|_\infty + \|h_\varepsilon\|_\infty)\, \varepsilon.$$

Moreover, we have $\|h\|^2, \|h_\varepsilon\|^2 \leq 1/\tau_0$. Then, there exists a constant $\mathbf{c} > 0$ such that

$$\begin{cases} \left| \|h\|_{n,N}^2 - \|h_\varepsilon\|_{n,N}^2 \right| \leq 2\sqrt{\frac{\mathbf{c}m}{\tau_0}}\varepsilon \ \text{ for the spline basis (see } \textit{Lemma 2.6} \text{ in Denis } \textit{et al.}(2021)) \\ \left| \|h\|_{n,N}^2 - \|h_\varepsilon\|_{n,N}^2 \right| \leq 2\sqrt{\frac{\mathbf{c}m}{\tau_0}}\varepsilon \ \text{ for an orthonormal basis } \ (\|h\|_\infty^2 \leq (\max_{0 \leq \ell \leq m-1} \|\phi_\ell\|_\infty^2) m \|h\|^2). \end{cases}$$

Therefore, for all $\delta > 0$ and for both the spline basis and any orthonormal basis,

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}_m} \left| \|h\|_{n,N}^2 - 1 \right| \geq \delta \right) \leq \mathbb{P}\left(\sup_{h \in \mathcal{H}_m^\varepsilon} \left| \|h\|_{n,N}^2 - 1 \right| \geq \delta/2 \right) + \mathbb{1}_{4\varepsilon\sqrt{\frac{\mathbf{c}m}{\tau_0}} \geq \delta}.$$

We set $\delta = 1/2$ and we choose $\varepsilon > 0$ such that $4\varepsilon\sqrt{\frac{\mathbf{c}m}{\tau_0}} < 1/2$. Then, using the Hoeffding inequality, there exists a constant $c > 0$ depending on $\mathbf{c}$ and τ_0 such that

$$\mathbb{P}\left(\Omega^c_{n,N,m}\right) \leq 2\mathcal{N}_\infty(\varepsilon, \mathcal{H}_m) \exp\left(-c\frac{N}{m}\right) \tag{2.54}$$

where $\mathcal{N}_\infty(\varepsilon, \mathcal{H}_m)$ is the covering number of $\mathcal{H}_m$ satisfying:

$$\mathcal{N}_\infty(\varepsilon, \mathcal{H}_m) \leq \left(\kappa\frac{\sqrt{m}}{\varepsilon}\right)^m \tag{2.55}$$

where the constant $\kappa > 0$ depends on $c > 0$ (see [30], *Proof of Lemma D.1*). We set $\varepsilon = \frac{\kappa\sqrt{m^*}}{N}$ with $m^* = \max \mathcal{M}$ and we derive from Equations (2.54) and (2.55) that

$$\mathbb{P}\left(\Omega^c_{n,N,m}\right) \leq 2N^{m^*} \exp\left(-c\frac{N}{m^*}\right) = 2\exp\left(-c\frac{N}{m^*}\left(1 - \frac{m^{*2}\log(N)}{cN}\right)\right).$$

- If $n \geq N$, then $m \in \mathcal{M} = \left\{1, \ldots, \sqrt{N}/\log(Nn)\right\}$. Since $m^{*2}\log(N)/N \to 0$ as $N \to +\infty$, there exists a constant $C > 0$ such that
$$\mathbb{P}\left(\Omega^c_{n,N,m}\right) \leq 2\exp\left(-C\sqrt{N}\right).$$
- If $n \leq N$, then $m \in \mathcal{M} = \{1, \ldots, \sqrt{n}/\log(Nn)\}$, $m^{*2}\log(N)/N \leq \log(N)/\log^2(Nn) \to 0$ as $N, n \to \infty$, and
$$\mathbb{P}\left(\Omega^c_{n,N,m}\right) \leq 2\exp\left(-C\sqrt{n}\right).$$
- If $n \propto N$, then $m \in \mathcal{M} = \left\{1, \ldots, \sqrt{N}/\log(Nn)\right\}$. Since $m^{*2}\log(N)/N \to 0$ as $N \to +\infty$, there exists a constant $C > 0$ such that
$$\mathbb{P}\left(\Omega^c_{n,N,m}\right) \leq 2\exp\left(-C\sqrt{N}\right).$$

□

***Proof of Theorem** 2.4.1* . The proof of Theorem 2.4.1 extends the proof of Theorem 2.3.2 when N tends to infinity. Then, we deduce from Equation (2.38) that

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_m - \sigma^2_{|I}\right\|^2_{n,N} \mathbb{1}_{\Omega_{n,N,m}}\right] \leq 3\inf_{h\in\mathcal{S}_{m,L}}\left\|h - \sigma^2_{|I}\right\|^2_n + C\mathbb{E}\left(\sup_{h\in\mathcal{S}_m, \|h\|_n=1}\nu^2(h)\right) + C\Delta^2 \tag{2.56}$$

where $C > 0$ is a constant depending on σ_1, and $\nu = \nu_1 + \nu_2 + \nu_3$ with

$$\nu_i(h) = \frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1} h(X^j_{k\Delta})\zeta^{j,i}_{k\Delta}, \quad i = 1, 2, 3$$

and the $\zeta^{j,i}_{k\Delta}$'s are the error terms depending on each path X^j, $j = 1, \ldots, N$.

Upper bound of $\mathbb{E}\left(\sup_{h\in\mathcal{S}_m,\ \|h\|_n=1}\nu^2(h)\right)$

For all $h = \sum_{\ell=0}^{m-1} a_\ell\phi_\ell \in \mathcal{S}_m$ such that $\|h\|_n = 1$, we have $\|h\|^2 \leq \frac{1}{\tau_0}$ and the coordinate vector $a = (a_0, \cdots, a_{m-1})$ satisfies:

- $\|a\|_2^2 \leq CK \leq Cm$ for the spline basis (see [30], Lemma 2.6)
- $\|a\|_2^2 \leq 1/\tau_0$ for an orthonormal basis since $\|h\|^2 = \|a\|_2^2$.

Furthermore, using the Cauchy−Schwarz inequality, we have:

$$\nu^2(h) = \left(\sum_{\ell=0}^{m-1} a_\ell \nu\left(\phi_\ell\right)\right)^2 \leq \|a\|_2^2 \sum_{\ell=0}^{m-1} \nu^2\left(\phi_\ell\right).$$

Thus, for all $\ell \in [\![0, m-1]\!]$, $\nu = \nu_1 + \nu_2 + \nu_3$ and for all $i \in \{1,2,3\}$

$$\mathbb{E}\left[\nu_i^2\left(\phi_\ell\right)\right] = \frac{1}{Nn^2}\mathbb{E}\left[\left(\sum_{k=0}^{n-1} \phi_\ell\left(X_{k\Delta}^1\right)\zeta_{k\Delta}^{1,i}\right)^2\right].$$

We finally deduce from (2.42), (2.43) and (2.44) that there exists a constant $C > 0$ depending on σ_1 such that:

$$\mathbb{E}\left(\sup_{h\in\mathcal{S}_m,\, \|h\|_n=1} \nu^2(h)\right) \leq C\frac{m}{Nn}. \tag{2.57}$$

We deduce from (2.56) and (2.57) that there exists a constant $C > 0$ such that,

$$\mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_{n,N}^2 \mathbb{1}_{\Omega_{n,N,m}}\right] \leq 3 \inf_{h\in\mathcal{S}_{m,L}} \left\|\sigma_{|I}^2 - h\right\|_n^2 + C\left(\frac{m}{Nn} + \Delta^2\right). \tag{2.58}$$

Since we have $\left\|\widehat{\sigma}_m^2\right\|_\infty \leq \sqrt{mL}$, then for m and L large enough, $\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_\infty^2 \leq 2mL$. There exists a constant $C > 0$ such that for all $m \in \mathcal{M}$ and for m and L large enough,

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_{n,N}^2\right] &= \mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_{n,N}^2 \mathbb{1}_{\Omega_{n,N,m}}\right] + \mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_{n,N}^2 \mathbb{1}_{\Omega_{n,N,m}^c}\right] \\
&\leq \mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_{n,N}^2 \mathbb{1}_{\Omega_{n,N,m}}\right] + 2mL\mathbb{P}\left(\Omega_{n,N,m}^c\right).
\end{aligned}$$

Then, from Equation (2.58), Lemma 2.8.8 and for $m \in \mathcal{M} = \left\{1, \ldots, \sqrt{\min(n,N)}/\sqrt{\log(Nn)}\right\}$, we have:

$$\mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_{n,N}^2\right] \leq 3 \inf_{h\in\mathcal{S}_{m,L}} \left\|h - \sigma_{|I}^2\right\|_n^2 + C\left(\frac{m}{Nn} + mL\exp\left(-C\sqrt{\min(n,N)}\right) + \Delta^2\right)$$

where $C > 0$ is a constant. Recall that the empirical norms $\|.\|_{n,N}$ and $\|.\|_n$ are equivalent on $\Omega_{n,N,m}$, that is for all $h \in \mathcal{S}_m$, $\|h\|_n^2 \leq 2\|h\|_{n,N}^2$. Thus, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_n^2\right] &= \mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_n^2 \mathbb{1}_{\Omega_{n,N,m}}\right] + \mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_n^2 \mathbb{1}_{\Omega_{n,N,m}^c}\right] \\
&\leq \mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_n^2 \mathbb{1}_{\Omega_{n,N,m}}\right] + 2mL\mathbb{P}\left(\Omega_{n,N,m}^c\right).
\end{aligned}$$

For all $h \in \mathcal{S}_{m,L} \subset \mathcal{S}_m$, we have:

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_n^2 \mathbb{1}_{\Omega_{n,N,m}}\right] &\leq 2\mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - h\right\|_n^2 \mathbb{1}_{\Omega_{n,N,m}}\right] + 2\left\|h - \sigma_{|I}^2\right\|_n^2 \\
&\leq 4\mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - h\right\|_{n,N}^2 \mathbb{1}_{\Omega_{n,N,m}}\right] + 2\left\|h - \sigma_{|I}^2\right\|_n^2 \\
&\leq 8\mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_{n,N}^2\right] + 10\left\|h - \sigma_{|I}^2\right\|_n^2.
\end{aligned}$$

We finally conclude that

$$\mathbb{E}\left[\left\|\widehat{\sigma}_m^2 - \sigma_{|I}^2\right\|_n^2\right] \leq 34 \inf_{h\in\mathcal{S}_{m,L}} \left\|h - \sigma_{|I}^2\right\|_n^2 + C\left(\frac{m}{Nn} + mL\exp\left(-C\sqrt{\min(n,N)}\right) + \Delta^2\right).$$

□

Proof of Theorem 2.4.3

Proof. We consider the restriction $\sigma^2\mathbb{1}_{[-\log(N),\log(N)]}$ of σ^2 on the compact interval $[-\log(N),\log(N)]$ on which the spline basis is built. Then we have:

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L}-\sigma^2\right\|_n^2\right]=\mathbb{E}\left[\left\|(\widehat{\sigma}^2_{m,L}-\sigma^2)\mathbb{1}_{[-\log(N),\log(N)]}\right\|_n^2\right]+\mathbb{E}\left[\left\|(\widehat{\sigma}^2_{m,L}-\sigma^2)\mathbb{1}_{[-\log(N),\log(N)]^c}\right\|_n^2\right]$$

and from Proposition 2.8.2, Lemma 2.8.4 and for N large enough, there exists constants $c, C>0$ such that

$$\begin{aligned}\mathbb{E}\left[\left\|(\widehat{\sigma}^2_{m,L}-\sigma^2)\mathbb{1}_{[-\log(N),\log(N)]^c}\right\|_n^2\right]\leq&\frac{2L}{n}\sum_{k=0}^{n-1}\mathbb{P}\left(|X_{k\Delta}|>\log(N)\right)\leq 2L\sup_{t\in[0,1]}\mathbb{P}\left(|X_t|\geq\log(N)\right)\\ \leq&\frac{C}{\log(N)}\exp\left(-c\log^2(N)\right).\end{aligned}$$

We deduce that

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L}-\sigma^2\right\|_n^2\right]=\mathbb{E}\left[\left\|(\widehat{\sigma}^2_{m,L}-\sigma^2)\mathbb{1}_{[-\log(N),\log(N)]}\right\|_n^2\right]+\frac{C}{\log(N)}\exp\left(-c\log^2(N)\right). \tag{2.59}$$

It remains to upper-bound the first term on the right hand side of Equation (2.59).

Upper bound of $\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L}-\sigma^2\right\|_n^2\mathbb{1}_{[-\log(N),\log(N)]}\right]$**.** For all $h\in\mathcal{S}_{m,L}$, we obtain from Equation (2.10),

$$\gamma_{n,N}(\widehat{\sigma}^2_{m,L})-\gamma_{n,N}(\sigma^2)\leq\gamma_{n,N}(h)-\gamma_{n,N}(\sigma^2). \tag{2.60}$$

For all $h\in\mathcal{S}_{m,L}$,

$$\gamma_{n,N}(h)-\gamma_{n,N}(\sigma^2)=\left\|h-\sigma^2\right\|_{n,N}^2+2\nu_1(\sigma^2-h)+2\nu_2(\sigma^2-h)+2\nu_3(\sigma^2-h)+2\mu(\sigma^2-h)$$

where

$$\nu_i(h)=\frac{1}{nN}\sum_{j=1}^{N}\sum_{k=0}^{n-1}h(X^j_{k\Delta})\zeta^{j,i}_{k\Delta},\quad i\in\{1,2,3\},\quad \mu(h)=\frac{1}{nN}\sum_{j=1}^{N}\sum_{k=0}^{n-1}h(X^j_{k\Delta})R^j_{k\Delta}, \tag{2.61}$$

we deduce from Equation (2.60) that for all $h\in\mathcal{S}_{m,L}$,

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L}-\sigma^2\right\|_{n,N}^2\mathbb{1}_{[-\log(N),\log(N)]}\right]\leq\inf_{h\in\mathcal{S}_{m,L}}\|h-\sigma^2\|_n^2+2\sum_{i=1}^{3}\mathbb{E}\left[\nu_i(\widehat{\sigma}^2_{m,L}-h)\right]+2\mathbb{E}\left[\mu(\widehat{\sigma}^2_{m,L}-h)\right]. \tag{2.62}$$

For all $i\in\{1,2,3\}$ and for all $h\in\mathcal{S}_{m,L}$, one has

$$\mathbb{E}\left[\nu_i\left(\widehat{\sigma}^2_{m,L}-h\right)\right]\leq\sqrt{2mL}\sqrt{\sum_{\ell=0}^{m-1}\mathbb{E}\left[\nu_i^2(\phi_\ell)\right]}. \tag{2.63}$$

1. Upper bound of $\sum_{\ell=0}^{m-1}\mathbb{E}\left[\nu_1^2(\phi_\ell)\right]$
   According to Equation (2.61), we have

$$\forall\ell\in[\![0,m-1]\!],\ \nu_1(\phi_\ell)=\frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\phi_\ell(X^j_{k\Delta})\zeta^{j,1}_{k\Delta}$$

   where $\zeta^{j,1}_{k\Delta}=\frac{1}{\Delta}\left[\left(\int_{k\Delta}^{(k+1)\Delta}\sigma(X^j_s)dW^j_s\right)^2-\int_{k\Delta}^{(k+1)\Delta}\sigma^2(X^j_s)ds\right]$ is a martingale satisfying

$$\mathbb{E}\left[\zeta^{1,1}_{k\Delta}|\mathcal{F}_{k\Delta}\right]=0\ \text{ and }\ \mathbb{E}\left[\left(\zeta^{1,1}_{k\Delta}\right)^2|\mathcal{F}_{k\Delta}\right]\leq\frac{1}{\Delta^2}\mathbb{E}\left[\left(\int_{k\Delta}^{(k+1)\Delta}\sigma^2(X^1_s)ds\right)^2\right]\leq C\sigma_1^4$$

with $C > 0$ a constant, $W = W^1$ and $(\mathcal{F}_t)_{t\geq 0}$ the natural filtration of the martingale $(M_t)_{t\in[0,1]}$ given for all $t \in [0,1]$ by $M_t = \int_0^t \sigma(X_s^1)dW_s$. We derive that

$$\sum_{\ell=0}^{m-1} \mathbb{E}\left[\nu_1^2(\phi_\ell)\right] = \frac{1}{Nn^2}\sum_{\ell=0}^{m-1}\mathbb{E}\left[\left(\sum_{k=0}^{n-1}\phi_\ell(X_{k\Delta}^j)\zeta_{k\Delta}^{1,1}\right)^2\right] = \frac{1}{Nn^2}\mathbb{E}\left[\sum_{k=0}^{n-1}\sum_{\ell=0}^{m-1}\phi_\ell^2(X_{k\Delta}^1)\left(\zeta_{k\Delta}^{1,1}\right)^2\right].$$

For each $k \in [\![0, n-1]\!]$, we have

$$\begin{cases} \sum_{\ell=0}^{m-1}\phi_\ell^2(X_{k\Delta}^1) = \sum_{\ell=-M}^{K-1} B_\ell^2(X_{k\Delta}^1) = 1 \ \text{ for the spline basis} \\ \\ \sum_{\ell=0}^{m-1}\phi_\ell^2(X_{k\Delta}^1) \leq Cm \ \text{ For an orthonormal basis with } C = \max\limits_{0\leq\ell\leq m-1}\|\phi_\ell\|_\infty^2. \end{cases}$$

Finally, there exists a constant $C > 0$ such that

$$\sum_{\ell=0}^{m-1}\mathbb{E}\left[\nu_1^2(\phi_\ell)\right] \leq \begin{cases} \frac{C}{Nn} \ \text{ for the spline basis} \\ \\ C\frac{m}{Nn} \ \text{ for an orthonormal basis.} \end{cases}$$

2. Upper bound of $\sum_{\ell=0}^{m-1}\mathbb{E}\left[\nu_2^2(\phi_\ell)\right]$
For all $k \in [\![0, n-1]\!]$ and for all $s \in [0,1]$, set $\eta(s) = k\Delta$ if $s \in [k\Delta, (k+1)\Delta)$. We have:

$$\begin{aligned}\sum_{\ell=0}^{m-1}\mathbb{E}\left[\nu_2^2(\phi_\ell)\right] &= \frac{4}{N}\sum_{\ell=0}^{m-1}\mathbb{E}\left[\left(\sum_{k=0}^{n-1}\int_{k\Delta}^{(k+1)\Delta}\phi_\ell(X_{k\Delta}^1)((k+1)\Delta - s)\sigma'(X_s^1)\sigma^2(X_s^1)dW_s\right)^2\right] \\ &= \frac{4}{N}\sum_{\ell=0}^{m-1}\mathbb{E}\left[\left(\int_0^1\phi_\ell(X_{\eta(s)}^1)(\eta(s)+\Delta-s)\sigma'(X_s^1)\sigma^2(X_s^1)dW_s\right)^2\right].\end{aligned}$$

We conclude that

$$\sum_{\ell=0}^{m-1}\mathbb{E}\left[\nu_2^2(\phi_\ell)\right] \leq \begin{cases} \frac{C}{Nn^2} \ \text{ for the spline basis} \\ \\ C\frac{m}{Nn^2} \ \text{ for an orthonormal basis.} \end{cases}$$

where the constant $C > 0$ depends on the diffusion coefficient and the upper bound of the basis functions.

3. Upper bound of $\sum_{\ell=0}^{m-1}\mathbb{E}\left[\nu_2^3(\phi_\ell)\right]$
We have:

$$\begin{aligned}\sum_{\ell=0}^{m-1}\mathbb{E}\left[\nu_3^2(\phi_\ell)\right] &= \frac{4}{Nn^2}\sum_{\ell=0}^{m-1}\mathbb{E}\left[\left(\sum_{k=0}^{n-1}\int_{k\Delta}^{(k+1)\Delta}\phi_\ell(X_{k\Delta}^1)b(X_{k\Delta}^1)\sigma(X_s^1)dW_s\right)^2\right] \\ &= \frac{4}{Nn^2}\sum_{\ell=0}^{m-1}\mathbb{E}\left[\left(\int_0^1\phi_\ell(X_{\eta(s)}^1)b(X_{\eta(s)}^1)\sigma(X_s^1)dW_s\right)^2\right] \\ &\leq \frac{4}{Nn^2}\mathbb{E}\left[\int_0^1\sum_{\ell=0}^{m-1}\phi_\ell^2(X_{\eta(s)}^1)b(X_{\eta(s)}^1)\sigma^2(X_s^1)ds\right].\end{aligned}$$

Since for all $x \in \mathbb{R}$, $b(x) \leq C_0(1+x^2)$ and $\sup\limits_{t\in[0,1]}\mathbb{E}\left(|X_t|^2\right) < \infty$, there exists a constant $C > 0$ depending on the diffusion coefficient such that

$$\sum_{\ell=0}^{m-1}\mathbb{E}\left[\nu_3^2(\phi_\ell)\right] \leq \begin{cases} \frac{C}{Nn^2} \ \text{ for the spline basis} \\ \\ C\frac{m}{Nn^2} \ \text{ for an orthonormal basis.} \end{cases}$$

We finally deduce that from Equations (2.62) and (3.46) that for all $h \in \mathcal{S}_{m,L}$,

$$\begin{cases} \mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L} - \sigma^2\right\|^2_{n,N} \mathbb{1}_{[-\log(N),\log(N)]}\right] \leq \inf_{h\in\mathcal{S}_{m,L}} \|h-\sigma^2\|^2_n + C\sqrt{\frac{mL}{Nn}} + 2\mathbb{E}\left[\mu(\widehat{\sigma}^2_{m,L} - h)\right] & (1) \\ \mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L} - \sigma^2\right\|^2_{n,N} \mathbb{1}_{[-\log(N),\log(N)]}\right] \leq \inf_{h\in\mathcal{S}_{m,L}} \|h-\sigma^2\|^2_n + C\sqrt{\frac{m^2L}{Nn}} + 2\mathbb{E}\left[\mu(\widehat{\sigma}^2_{m,L} - h)\right] & (2) \end{cases} \tag{2.64}$$

where $C > 0$ is a constant, the result (1) corresponds to the spline basis, and the result (2) corresponds to any orthonormal basis. It remains to obtain an upper bound of the term $\mu(\widehat{\sigma}^2_{m,L} - h)$. For all $a > 0$ and for all $h \in \mathcal{S}_{m,L}$,

$$\begin{aligned} 2\mu\left(\widehat{\sigma}^2_{m,L} - h\right) \leq & \frac{2}{a}\left\|\widehat{\sigma}^2_{m,L} - \sigma^2\right\|^2_{n,N} + \frac{2}{a}\left\|h-\sigma^2\right\|^2_{n,N} + \frac{a}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\left(R^j_{k\Delta}\right)^2 \\ 2\mathbb{E}\left[\mu\left(\widehat{\sigma}^2_{m,L} - h\right)\right] \leq & \frac{2}{a}\mathbb{E}\left\|\widehat{\sigma}^2_{m,L} - \sigma^2\right\|^2_{n,N} + \frac{2}{a}\inf_{h\in\mathcal{S}_{m,L}}\|h-\sigma^2\|^2_n + \frac{a}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\mathbb{E}\left[\left(R^j_{k\Delta}\right)^2\right]. \end{aligned}$$

Using Equations (2.37), (2.64) and setting $a = 4$, we deduce that there exists a constant $C > 0$ depending on σ_1 such that,

$$\begin{cases} \mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L} - \sigma^2\right\|^2_{n,N} \mathbb{1}_{[-\log(N),\log(N)]}\right] \leq \inf_{h\in\mathcal{S}_{m,L}} \|h-\sigma^2\|^2_n + C\left(\sqrt{\frac{mL}{Nn}} + \Delta^2\right) & [\mathbf{B}] \\ \mathbb{E}\left[\left\|\widehat{\sigma}^2_{m,L} - \sigma^2\right\|^2_{n,N} \mathbb{1}_{[-\log(N),\log(N)]}\right] \leq \inf_{h\in\mathcal{S}_{m,L}} \|h-\sigma^2\|^2_n + C\left(\sqrt{\frac{m^2L}{Nn}} + \Delta^2\right) & [\mathbf{H}]. \end{cases} \tag{2.65}$$

The final result is obtained from Equations (2.59) and (2.65). □

Proof of Lemma 2.4.5

Proof. It is proven in [18] that for each dimension $m \in \mathcal{M}$, the Gram matrix Ψ_m built from the Hermite basis is invertible. For the case of the **B**-spline basis, let us consider a vector $(x_{-M}, \cdots, x_{K-1}) \in \mathbb{R}^m$ such that $x_j \in [u_{j+M}, u_{j+M+1})$ and $B_j(x_j) \neq 0$. Since $[u_{j+M}, u_{j+M+1}) \cap [u_{j'+M}, u_{j'+M+1}) = \emptyset$ for all $j, j' \in \{-M, \cdots, K-1\}$ such that $j \neq j'$, then for all $j, j' \in \{-M, \cdots, K-1\}$ such that $j \neq j', B_j(x_{j'}) = 0$. Consequently, we obtain:

$$\begin{aligned} \det\left((B_\ell(x_{\ell'}))_{-M\leq\ell,\ell'\leq K-1}\right) &= \det\left(\operatorname{diag}\left(B_{-M}(x_M), \cdots, B_{K-1}(x_{K-1})\right)\right) \\ &= \prod_{\ell=-M}^{K-1} B_\ell(x_\ell) \neq 0. \end{aligned}$$

Then, we deduce from [18], *Lemma 1* that the matrix Ψ_m is invertible for all $m \in \mathcal{M}$, where the function f_T are replaced by

$$f_n : x \mapsto \frac{1}{n}\sum_{k=0}^{n-1} p_X(k\Delta, x)$$

with $\lambda([-A_N, A_N] \cap \operatorname{supp}(f_n)) > 0$, λ being the Lebesgue measure.

Case of the B-spline basis. For all $w \in \mathbb{R}^m$ such that $\|w\|_{2,m} = 1$, we have:

$$w'\Psi_m w = \|t_w\|^2_n = \int_{-A_N}^{A_N} t^2_w(x) f_n(x)dx + \frac{t^2_w(x_0)}{n} \text{ with } t_w = \sum_{\ell=-M}^{K-1} w_\ell B_\ell.$$

Under Assumption 2.2.1, the transition density $(t,x) \mapsto p_X(t,x)$ is approximated as follows

$$\forall (t,x) \in (0,1] \times \mathbb{R}, \quad \frac{1}{K_* \sqrt{t}} \exp\left(-c_\sigma \frac{x^2}{t}\right) \leq p_X(t,x) \leq \frac{K_*}{\sqrt{t}} \exp\left(-\frac{x^2}{c_\sigma t}\right)$$

where $K_* > 1$ and $c_\sigma > 1$. Since $s \mapsto \exp\left(-c_\sigma x^2/s\right)$ is an increasing function, then for n large enough and for all $x \in [-A_N, A_N]$,

$$\begin{aligned} f_n(x) &\geq \frac{1}{K_* n} \sum_{k=1}^{n-1} \exp\left(-c \frac{x^2}{k\Delta}\right) \geq \frac{1}{K_*} \int_0^{1-\Delta} \exp\left(-c_\sigma \frac{x^2}{s}\right) ds \\ &\geq \frac{1}{K_*} \int_{1-(\log(N))^{-1}}^{1-(2\log(N))^{-1}} \exp\left(-c_\sigma \frac{x^2}{s}\right) ds \\ &\geq \frac{1}{2K_* \log(N)} \exp\left(-\frac{c_\sigma x^2}{1-\log^{-1}(N)}\right). \end{aligned}$$

Thus, the density function satisfies

$$\forall x \in [-A_N, A_N], \quad f_n(x) \geq \frac{1}{2K_* \log(N)} \exp\left(-\frac{c_\sigma A_N^2}{1-\log^{-1}(N)}\right) \geq \frac{1}{2K_* \log(N)} \exp\left(-c_\sigma A_N^2\right). \quad (2.66)$$

Finally, since there exists a constant $C_1 > 0$ such that $\|t_w\|^2 \geq C_1 A_N K_N^{-1}$ (see [30], Lemma 2.6), for all $w \in \mathbb{R}^m$ ($m = K_N + M$) such that $\|w\|_{2,m} = 1$, there exists a constant $C > 0$ such that,

$$w' \Psi_m w \geq \frac{C A_N}{m \log(N)} \exp\left(-c_\sigma A_N^2\right).$$

Case of the Hermite basis. For all $w \in \mathbb{R}^m$ such that $\|w\|_{2,m} = 1$, we have

$$w' \Psi_m w = \|t_w\|_n^2 = \int_{-\infty}^{+\infty} t_w^2(x) f_n(x) dx + \frac{t_w^2(x_0)}{n} \quad \text{with} \quad t_w = \sum_{\ell=0}^{m-1} w_\ell h_\ell.$$

Recall that for all $x \in \mathbb{R}$ such that $|x| \geq \sqrt{(3/2)(4m+3)}$, $|h_\ell(x)| \leq c|x| \exp(-c_0 x^2)$ for all $\ell \geq 0$. Then we have

$$\begin{aligned} w' \Psi_m w &\geq \int_{|x| \leq \sqrt{(3/2)(4m+3)}} \left(\sum_{\ell=0}^{m-1} w_\ell h_\ell(x)\right)^2 f_n(x) dx \\ &\geq \inf_{x \in [-\sqrt{(3/2)(4m+3)}, \sqrt{(3/2)(4m+3)}]} f_n(x) \int_{|x| \leq \sqrt{(3/2)(4m+3)}} \left(\sum_{\ell=0}^{m-1} w_\ell h_\ell(x)\right)^2 dx \\ &\geq \frac{1}{2K_* \log(N)} \exp\left(-\frac{3c_\sigma(4m+3)}{2\left(1-\log^{-1}(N)\right)}\right) \int_{|x| \leq \sqrt{(3/2)(4m+3)}} \left(\sum_{\ell=0}^{m-1} w_\ell h_\ell(x)\right)^2 dx \end{aligned}$$

since for all $x \in \mathbb{R}$, $f_n(x) \geq (1/2K_* \log(N)) \exp\left(-\frac{c_\sigma x^2}{1-\log^{-1}(N)}\right)$. Set $a_N = \sqrt{(3/2)(4m+3)}$, then we obtain

$$\begin{aligned} w' \Psi_m w &\geq \frac{\exp\left(-\frac{3c_\sigma(4m+3)}{2\left(1-\log^{-1}(N)\right)}\right)}{2K_* \log(N)} \left(\int_{-\infty}^{+\infty} \left(\sum_{\ell=0}^{m-1} w_\ell h_\ell(x)\right)^2 dx - \int_{|x|>a_N} \left(\sum_{\ell=0}^{m-1} w_\ell h_\ell(x)\right)^2 dx\right) \\ &\geq \frac{\exp\left(-\frac{3c_\sigma(4m+3)}{2\left(1-\log^{-1}(N)\right)}\right)}{2K_* \log(N)} \left(1 - 2c^2 m \int_{a_N}^{+\infty} x^2 \exp(-8c_0 x^2) dx\right) \\ &\geq \frac{\exp\left(-\frac{3c_\sigma(4m+3)}{2\left(1-\log^{-1}(N)\right)}\right)}{2K_* \log(N)} \left(1 - \frac{c^2 m}{8c_0} \sqrt{\frac{3}{2}(4m+3)} \exp\left(-12c_0(4m+3)\right)\right) \end{aligned}$$

where $c, c_0 > 0$ are constants depending on the Hermite basis. Finally, for N large enough,

$$1 - \frac{c^2 m}{8c_0}\sqrt{\frac{3}{2}(4m+3)}\exp\left(-12c_0(4m+3)\right) \geq \frac{1}{2}.$$

Finally, there exists a constant $C > 0$ such that for all $w \in \mathbb{R}^m$ such that $\|w\|_{2,m}$,

$$w'\Psi_m w \geq \frac{C}{\log(N)}\exp\left(-\frac{3c_\sigma(4m+3)}{2\left(1-\log^{-1}(N)\right)}\right).$$

□

Proof of Theorem 2.4.6

The proof of Theorem 2.4.6 relies on the following lemma:

Lemma 2.8.9. *Under Assumptions 2.2.1 and for $\sigma^2 \in \Sigma_I(\beta, R)$ with $\beta \geq 1$, $I = [-A_N, A_N]$ and*

$$N \propto n,\ A_N = o\left(\sqrt{\log(N)}\right),\ K \propto \left((Nn)^{1/(2\beta+1)}A_N\right)\ \ (m = K + M),$$

the following holds:

$$\mathbb{P}\left(\Omega^c_{n,N,m}\right) \leq C\exp\left(-c\log^{3/2}(N)\right)$$

where $c, C > 0$ are constants independent of N.

Proof of Theorem 2.4.6 . According to Equations (2.56) in the proof of Theorem 2.4.1, for all dimension $m = K + M$, with $K \in \mathcal{M}$, and for all $h \in \mathcal{S}_{K+M}$, there exists a constant $C > 0$ such that

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{A_N,m} - \sigma^2_{A_N}\right\|^2_{n,N}\mathbb{1}_{\Omega_{n,N,m}}\right] \leq C\left[\inf_{h\in\mathcal{S}_{K+M,L}}\|h - \sigma^2_{A_N}\|^2_n + \mathbb{E}\left(\sup_{h\in\mathcal{S}_{K+M},\|h\|_n=1}\nu^2(h)\right) + \Delta^2\right] \quad (2.67)$$

where $\Omega_{n,N,m}$ is given in Equation (2.53) and $\nu = \nu_1 + \nu_2 + \nu_3$ with the ν_i given in Equation (2.33) .

For all $h = \sum_{\ell=-M}^{K-1} a_\ell B_\ell \in \mathcal{S}_{K+M,L_N}$,

$$\|h\|^2_n = \mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1}h^2(X_{k\Delta})\right] = \sum_{\ell=-M}^{K-1}\sum_{\ell=-M}^{K-1}a_\ell a_{\ell'}\mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1}B_\ell(X_{k\Delta})B_{\ell'}(X_{k\Delta})\right] = a'\Psi_m a.$$

The Gram matrix Ψ_m is invertible for each $K \in \mathcal{M}$ (see proof of Lemma 2.4.5). It follows that for all $h = \sum_{\ell=-M}^{K-1} a_\ell B_\ell$ such that $\|h\|^2_n = a'\Psi_m a = 1$, one has $a = \Psi_m^{-1/2}u$ where $u \in \mathbb{R}^m$ and $\|u\|_{2,m} = 1$. Furthermore, we have:

$$h = \sum_{\ell=-M}^{K-1}a_\ell B_\ell = \sum_{\ell=-M}^{K-1}u_\ell\sum_{\ell'=-M}^{K-1}\left[\Psi_m^{-1/2}\right]_{\ell',\ell}B_{\ell'}.$$

Then for all $h \in \mathcal{S}_{K+M}$, we have $\nu^2(h) \leq 3(\nu_1^2(h) + \nu_2^2(h) + \nu_3^2(h))$ where,

$$\forall i \in \{1,2,3\},\ \nu_i^2(h) \leq \sum_{\ell=-M}^{K-1}\left(\frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\sum_{\ell'=-M}^{K-1}\left[\Psi_m^{-1/2}\right]_{\ell',\ell}B_{\ell'}(X^j_{k\Delta})\zeta^{j,i}_{k\Delta}\right)^2.$$

So we obtain,

$$\forall i \in \{1,2,3\},\ \mathbb{E}\left[\sup_{h\in\mathcal{S}_{K+M},\|h\|_n=1}\nu_i^2(h)\right] \leq \frac{1}{Nn^2}\sum_{\ell=-M}^{K-1}\mathbb{E}\left[\left(\sum_{k=0}^{n-1}\sum_{\ell'=-M}^{K-1}\left[\Psi_m^{-1/2}\right]_{\ell',\ell}B_{\ell'}(X^j_{k\Delta})\zeta^{1,i}_{k\Delta}\right)^2\right]$$

For $i=1$, we have $\zeta_{k\Delta}^{1,1}=\frac{1}{\Delta}\left[\left(\int_{k\Delta}^{(k+1)\Delta}\sigma(X_s^1)dW_s\right)^2-\int_{k\Delta}^{(k+1)\Delta}\sigma^2(X_s^1)ds\right]$ and we obtained in the proof of Theorem 2.5.1 that there exists a constant $C>0$ such that for all $k\in[\![0,n-1]\!]$,

$$\mathbb{E}\left[\zeta_{k\Delta}^{1,1}|\mathcal{F}_{k\Delta}\right]=0,\quad \mathbb{E}\left[\left(\zeta_{k\Delta}^{1,1}\right)^2|\mathcal{F}_{k\Delta}\right]\leq C\mathbb{E}\left[\left(\int_{k\Delta}^{(k+1)\Delta}\sigma^2(X_u)du\right)^2\right]\leq C\sigma_1^4\Delta^2.$$

We deduce that

$$\begin{aligned}\mathbb{E}\left[\sup_{h\in\mathcal{S}_{K+M},\|h\|_n=1}\nu_1^2(h)\right]&=\frac{1}{Nn^2\Delta^2}\sum_{\ell=0}^{K-1}\mathbb{E}\left[\left(\sum_{k=0}^{n-1}\sum_{\ell'=-M}^{K-1}\left[\Psi_m^{-1/2}\right]_{\ell',\ell}B_{\ell'}(X_{k\Delta}^1)\zeta_{k\Delta}^{1,1}\right)^2\right]\\&\leq\frac{1}{N}\sum_{\ell=-M}^{K-1}\sum_{k=0}^{n-1}\mathbb{E}\left\{\left(\sum_{\ell'=-M}^{K-1}\left[\Psi_m^{-1/2}\right]_{\ell',\ell}B_{\ell'}(X_{k\Delta}^1)\right)^2\left(\zeta_{k\Delta}^{1,1}\right)^2\right\}\\&\leq\frac{4\sigma_1^2}{Nn}\sum_{\ell=-M}^{K-1}\sum_{\ell'=-M}^{K-1}\sum_{\ell''=-M}^{K-1}\left[\Psi_m^{-1/2}\right]_{\ell',\ell}\left[\Psi_m^{-1/2}\right]_{\ell'',\ell}\left[\Psi_m^{-1/2}\right]_{\ell',\ell''}.\end{aligned}$$

We have:

$$\sum_{\ell=-M}^{K-1}\sum_{\ell'=-M}^{K-1}\sum_{\ell''=-M}^{K-1}\left[\Psi_m^{-1/2}\right]_{\ell',\ell}\left[\Psi_m^{-1/2}\right]_{\ell'',\ell}\left[\Psi_m^{-1/2}\right]_{\ell',\ell''}=\mathrm{Tr}\left(\Psi_m^{-1}\Psi_m\right)=\mathrm{Tr}(I_m)=m.$$

So we obtain

$$\mathbb{E}\left[\sup_{h\in\mathcal{S}_{K+M}}\nu_1^2(h)\right]\leq\frac{4\sigma_1^2m}{Nn}.$$

For $i=2$, we have $\zeta_{k\Delta}^{1,2}=\frac{2}{\Delta}\int_{k\Delta}^{(k+1)\Delta}\left((k+1)\Delta-s\right)\sigma'(X_s^1)\sigma^2(X_s^1)dW_s$ and

$$\begin{aligned}\mathbb{E}\left[\sup_{h\in\mathcal{S}_{K+M},\|h\|_n=1}\nu_2^2(h)\right]&\leq\frac{1}{Nn^2}\sum_{\ell=-M}^{K-1}\mathbb{E}\left[\left(\sum_{k=0}^{n-1}\sum_{\ell'=-M}^{K-1}\left[\Psi_m^{-1/2}\right]_{\ell',\ell}B_{\ell'}(X_{k\Delta}^j)\zeta_{k\Delta}^{1,2}\right)^2\right]\\&\leq\frac{4\sigma_1^4\|\sigma'\|_\infty^2\Delta}{Nn^2}\sum_{\ell=-M}^{K-1}\sum_{k=0}^{n-1}\mathbb{E}\left[\left(\sum_{\ell'=-M}^{K-1}\left[\Psi_m^{-1/2}\right]_{\ell',\ell}B_{\ell'}(X_{k\Delta}^1)\right)^2\right]\\&\leq\frac{4\sigma_1^2\|\sigma'\|_\infty^2m}{Nn^2}.\end{aligned}$$

For $i=3$, we have $\zeta_{k\Delta}^{1,3}=2b\left(X_{k\Delta}^1\right)\int_{k\Delta}^{(k+1)\Delta}\sigma(X_s^1)dW_s$ and there exists constants $C_1,C_2>0$ such that

$$\begin{aligned}\mathbb{E}\left[\sup_{h\in\mathcal{S}_{K+M},\|h\|_n=1}\nu_3^2(h)\right]&\leq\frac{1}{Nn^2}\sum_{\ell=-M}^{K-1}\mathbb{E}\left[\left(\sum_{k=0}^{n-1}\sum_{\ell'=-M}^{K-1}\left[\Psi_m^{-1/2}\right]_{\ell',\ell}B_{\ell'}(X_{k\Delta}^j)\zeta_{k\Delta}^{1,3}\right)^2\right]\\&\leq C_1\frac{\sigma_1^2\Delta}{Nn^2}\sum_{\ell=-M}^{K-1}\sum_{k=0}^{n-1}\mathbb{E}\left[\left(\sum_{\ell'=-M}^{K-1}\left[\Psi_m^{-1/2}\right]_{\ell',\ell}B_{\ell'}(X_{k\Delta}^1)\right)^2\right]\\&\leq C_2\frac{\sigma_1^2m}{Nn^2}.\end{aligned}$$

Finally, there exists a constant $C>0$ depending on σ_1 and M such that

$$\mathbb{E}\left[\sup_{h\in\mathcal{S}_{K+M},\|h\|_n=1}\nu^2(h)\right]\leq C\frac{m}{Nn}.\tag{2.68}$$

From Equations (2.67) and (2.68) , we deduce that

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{A_N,m}-\sigma^2_{A_N}\right\|^2_{n,N}\mathbb{1}_{\Omega_{n,N,m}}\right]\leq C\left(\inf_{h\in\mathcal{S}_{K+M,L}}\|h-\sigma^2_{A_N}\|^2_n+\frac{m}{Nn}+\Delta^2\right)$$

where $C>0$ is a constant depending on σ_1 and M. We obtain

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{A_N,m}-\sigma^2_{A_N}\right\|^2_{n,N}\right]\leq C\left(\inf_{h\in\mathcal{S}_{K+M,L}}\|h-\sigma^2_{A_N}\|^2_n+\frac{m}{Nn}+\Delta^2\right)+\mathbb{E}\left[\left\|\widehat{\sigma}^2_{A_N,m}-\sigma^2_{A_N}\right\|^2_{n,N}\mathbb{1}_{\Omega^c_{n,N,m}}\right]$$

and for N large enough, $\left\|\widehat{\sigma}^2_{A_N,m}-\sigma^2_{A_N}\right\|^2_{n,N}\leq 4mL$, and according to Lemma 2.8.9 ,

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{A_N,m}-\sigma^2_{A_N}\right\|^2_{n,N}\mathbb{1}_{\Omega^c_{n,N,m}}\right]\leq 4mL\mathbb{P}\left(\Omega^c_{n,N,m}\right)\leq CmL\exp\left(-c\log^{3/2}(N)\right)$$

where $c>0$ is a constant. Thus, there exists a constant $C>0$ such that

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\sigma}^2_{A_N,m}-\sigma^2_{A_N}\right\|^2_{n,N}\right]&\leq\mathbb{E}\left[\left\|\widehat{\sigma}^2_{A_N,m}-\sigma^2_{A_N}\right\|^2_{n,N}\mathbb{1}_{\Omega_{n,N,m}}\right]+\mathbb{E}\left[\left\|\widehat{\sigma}^2_{A_N,m}-\sigma^2_{A_N}\right\|^2_{n,N}\mathbb{1}_{\Omega^c_{n,N,m}}\right]\\&\leq C\left(\inf_{h\in\mathcal{S}_{K+M,L}}\|h-\sigma^2_{A_N}\|^2_n+\frac{m}{Nn}+mL\exp\left(-c\log^{3/2}(N)\right)+\Delta^2\right).\end{aligned}$$

Then, as $n\propto N$ and $L=\log(N)$, there exists a constant $C>0$ such that

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{A_N,m}-\sigma^2_{A_N}\right\|^2_{n,N}\right]\leq C\left(\inf_{h\in\mathcal{S}_{K+M,L}}\|h-\sigma^2_{A_N}\|^2_n+\frac{m}{Nn}\right).$$

Finally, since $\sigma^2\in\Sigma_I(\beta,R)$ with $\beta\geq 1$ and $I=[-A_N,A_N]$, one has

$$\inf_{h\in\mathcal{S}_{K+M,L}}\|h-\sigma^2_{A_N}\|^2_n\leq CA_N^{2\beta}K^{-2\beta}$$

where the constant $C>0$ depends on β, R and M. Furthermore, as we chose the inverval $[-A_N,A_N]$ such that $A_N=o\left(\sqrt{\log(N)}\right)$ and for $K\propto\left((Nn)^{1/(2\beta+1)}A_N\right)$, we obtain

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{A_N,m}-\sigma^2_{A_N}\right\|^2_{n,N}\right]\leq C\log^{\beta}(N)(Nn)^{-2\beta/(2\beta+1)}.$$

□

2.8.4 Proof of Section 2.5

Proof of Theorem 2.5.1

Set for all $K,K'\in\mathcal{K}=\left\{2^q,\ q=0,\dots,q_{\max},\ 2^{q_{\max}}\leq\sqrt{N}/\log(N)\right\}\subset\mathcal{M}$,

$$\mathcal{T}_{K,K'}=\left\{g\in\mathcal{S}_{K+M}+\mathcal{S}_{K'+M},\ \|g\|_n=1,\ \|g\|_\infty\leq\sqrt{L}\right\}. \tag{2.69}$$

Recall that for all $j\in[\![1,N]\!]$ and for all $k\in[\![0,n]\!]$,

$$\zeta^{j,1}_{k\Delta}=\frac{1}{\Delta}\left[\left(\int_{k\Delta}^{(k+1)\Delta}\sigma(X^j_s)dW^j_s\right)^2-\int_{k\Delta}^{(k+1)\Delta}\sigma^2(X^j_s)ds\right].$$

The proof of Theorem 2.5.1 relies on the following lemma whose proof is in Appendix.

Lemma 2.8.10. *Under Assumption 2.2.1, for all $\varepsilon,\ v>0$ and $g\in\mathcal{T}_{K,K'}$, there exists a real constant $C>0$ such that,*

$$\mathbb{P}\left(\frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}g(X^j_{k\Delta})\zeta^{j,1}_{k\Delta}\geq\varepsilon,\|g\|^2_{n,N}\leq v^2\right)\leq\exp\left(-C\frac{Nn\varepsilon^2}{\sigma_1^2\left(\varepsilon\|g\|_\infty+4\sigma_1^2v^2\right)}\right)$$

and for all $x>0$ such that $x\leq\varepsilon^2/\sigma_1^2\left(\varepsilon\|g\|_\infty+4\sigma_1^2v^2\right)$,

$$\mathbb{P}\left(\frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}g(X^j_{k\Delta})\zeta^{j,1}_{k\Delta}\geq 2\sigma_1^2v\sqrt{x}+\sigma_1^2\|g\|_\infty x,\|g\|^2_{n,N}\leq v^2\right)\leq\exp\left(-CNnx\right).$$

Proof of Theorem 2.5.1. From Equation (2.23), we have

$$\widehat{K} := \underset{K\in\mathcal{K}}{\arg\min}\{\gamma_{n,N}(\widehat{\sigma}^2_K) + \text{pen}(K)\}.$$

For all $K \in \mathcal{K}$ and $h \in \mathcal{S}_{K+M,L}$,

$$\gamma_{n,N}(\widehat{\sigma}^2_{\widehat{K}}) + \text{pen}(\widehat{K}) \leq \gamma_{n,N}(h) + \text{pen}(K),$$

then, for all $K \in \mathcal{K}$ and for all $h \in \mathcal{S}_{K+M,L}$,

$$\begin{aligned}
\gamma_{n,N}(\widehat{\sigma}^2_{\widehat{K}}) - \gamma_{n,N}(\sigma^2_{|I}) &\leq \gamma_{n,N}(h) - \gamma_{n,N}(\sigma^2_{|I}) + \text{pen}(K) - \text{pen}(\widehat{K}) \\
\left\|\widehat{\sigma}^2_{\widehat{K}} - \sigma^2_{|I}\right\|^2_{n,N} &\leq \left\|h - \sigma^2_{|I}\right\|^2_{n,N} + 2\nu\left(\widehat{\sigma}^2_{\widehat{K}} - h\right) + 2\mu\left(\widehat{\sigma}^2_{\widehat{K}} - h\right) + \text{pen}(K) - \text{pen}(\widehat{K}) \\
&\leq \left\|h - \sigma^2_{|I}\right\|^2_{n,N} + \frac{1}{d}\left\|\widehat{\sigma}^2_{\widehat{K}} - t\right\|^2_n + d\sup_{g\in\mathcal{T}_{K,\widehat{K}}}\nu^2(g) + \frac{1}{d}\left\|\widehat{\sigma}^2_{\widehat{K}} - h\right\|^2_{n,N} \\
&\quad + \frac{d}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\left(R^j_{k\Delta}\right)^2 + \text{pen}(K) - \text{pen}(\widehat{K})
\end{aligned}$$

where $d > 1$ and the space $\mathcal{T}_{K,\widehat{K}}$ is given in Equation (2.69). On the set $\Omega_{n,N,K_{\max}}$ (given in Equation (2.53)): $\forall h \in \mathcal{S}_{K+M}$, $\frac{1}{2}\|h\|^2_n \leq \|h\|^2_{n,N} \leq \frac{3}{2}\|h\|^2_n$. Then on $\Omega_{n,N,K_{\max}}$, for all $d > 1$ and for all $h \in \mathcal{S}_{K+M}$ with $K \in \mathcal{K}$,

$$\begin{aligned}
\left(1 - \frac{10}{d}\right)\left\|\widehat{\sigma}^2_{\widehat{K}} - \sigma^2_{|I}\right\|^2_{n,N} \leq \left(1 + \frac{10}{d}\right)\left\|h - \sigma^2_{|I}\right\|^2_{n,N} + d\sup_{h\in\mathcal{T}_{K,\widehat{K}}}\nu^2(h) + \frac{d}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\left(R^j_{k\Delta}\right)^2 \\
+ \text{pen}(K) - \text{pen}(\widehat{K}).
\end{aligned}$$

We set $d = 20$. Then, on $\Omega_{n,N,\max}$ and for all $h \in \mathcal{S}_{K+M,L}$,

$$\left\|\widehat{\sigma}^2_{\widehat{K}} - \sigma^2_{|I}\right\|^2_{n,N} \leq 3\left\|h - \sigma^2_{|I}\right\|^2_{n,N} + 20\sup_{h\in\mathcal{T}_{K,\widehat{K}}}\nu^2(h) + \frac{20}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\left(R^j_{k\Delta}\right)^2 + 2\left(\text{pen}(K) - \text{pen}(\widehat{K})\right). \tag{2.70}$$

Let $q : \mathcal{K}^2 \longrightarrow \mathbb{R}_+$ such that $160q(K,K') \leq 18\text{pen}(K) + 16\text{pen}(K')$. Thus, on the set $\Omega_{n,N,K_{\max}}$, there exists a constant $C > 0$ such that for all $h \in \mathcal{S}_{K+M}$

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\sigma}^2_{\widehat{K}} - \sigma^2_{|I}\right\|^2_{n,N}\mathbb{1}_{\Omega_{n,N,K_{\max}}}\right] &\leq 34\left(\inf_{h\in\mathcal{S}_{K+M,L}}\left\|h - \sigma^2_{|I}\right\|^2_n + \text{pen}(K)\right) \\
&\quad + 160\mathbb{E}\left(\sup_{h\in\mathcal{T}_{K,\widehat{K}}}\nu^2_1(h) - q(K,\widehat{K})\right) + C\Delta^2
\end{aligned}$$

where

$$\nu_1(h) := \frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}h(X^j_{k\Delta})\zeta^{j,1}_{k\Delta} \tag{2.71}$$

with $\zeta^{j,1}_{k\Delta}$ the error term. We set for all $K, K' \in \mathcal{K}$,

$$G_K(K') := \sup_{h\in\mathcal{T}_{K,K'}}\nu^2_1(h) \tag{2.72}$$

and for N and n large enough, $\left\|\widehat{\sigma}^2_{\widehat{K}} - \sigma^2_{|I}\right\|^2_{n,N} \leq 4(K+M)L$. We deduce that,

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\sigma}^2_{\widehat{K}} - \sigma^2_{|I}\right\|^2_{n,N}\right] &\leq \mathbb{E}\left[\left\|\widehat{\sigma}^2_{\widehat{K}} - \sigma^2_{|I}\right\|^2_{n,N}\mathbb{1}_{\Omega_{n,N,K_{\max}}}\right] + \mathbb{E}\left[\left\|\widehat{\sigma}^2_{\widehat{K}} - \sigma^2_{|I}\right\|^2_{n,N}\mathbb{1}_{\Omega^c_{n,N,K_{\max}}}\right] \\
&\leq 34\inf_{K\in\mathcal{K}}\left(\inf_{h\in\mathcal{S}_{K+M}}\left\|h - \sigma^2_{|I}\right\|^2_n + \text{pen}(K)\right) \\
&\quad + C\Delta^2 + 4(K+M)L\mathbb{P}(\Omega^c_{n,N,K_{\max}}) \\
&\quad + 160\mathbb{E}\left[\left(G_K(\widehat{K}) - q(K,\widehat{K})\right)_+\mathbb{1}_{\Omega_{n,N,K_{\max}}}\right].
\end{aligned}$$

Let $\overline{B}_{\|.\|_n}(0,1)$ be the unit ball of the approximation subspace $\mathcal{S}_{K+M}$ with respect to norm $\|.\|_n$, defined as follows:

$$\overline{B}_{\|.\|_n}(0,1) = \{h \in \mathcal{S}_{K+M} : \|h\|_n \leq 1\} = \left\{h \in \mathcal{S}_{K+M} : \|h\| \leq \frac{1}{\tau_0}\right\} = \overline{B}_2(0, 1/\tau_0),$$

we can find a $\varepsilon-$net E_ε such that for each $\varepsilon \in (0,1]$, $|E_\varepsilon| \leq \left(\frac{3}{\varepsilon\tau_0}\right)^{K+M}$ (see [69]).

Recall that $\mathcal{T}_{K,K'} = \left\{g \in \mathcal{S}_{K+M} + \mathcal{S}_{K'+M}, \|g\|_n = 1, \|g\|_\infty \leq \sqrt{L}\right\}$ and consider the sequence $(E_{\varepsilon_k})_{k\geq 1}$ of $\varepsilon-$net with $\varepsilon_k = \varepsilon_0 2^{-k}$ and $\varepsilon_0 \in (0,1]$. Moreover, set $N_k = \log(|E_{\varepsilon_k}|)$ for each $k \geq 0$. Then for each $g \in \mathcal{S}_{K+M} + \mathcal{S}_{K'+M}$ such that $\|g\|_\infty \leq \sqrt{L}$, there exists a sequence $(g_k)_{k\geq 0}$ with $g_k \in E_{\varepsilon_k}$ such that $g = g_0 + \sum_{k=1}^{\infty} g_k - g_{k-1}$. Set $\widetilde{\mathbb{P}} := \mathbb{P}\left(. \cap \Omega_{n,N,K_{\max}}\right)$ and

$$\tau := \sigma_1^2\sqrt{6x_0^{n,N}} + \sigma_1^2\sqrt{L}x_0^{n,N} + \sum_{k\geq 1}\varepsilon_{k-1}\left\{\sigma_1^2\sqrt{6x_k^{n,N}} + 2\sigma_1^2\sqrt{L}x_k^{n,N}\right\} = y_0^{n,N} + \sum_{k\geq 0} y_k^{n,N}.$$

For all $h \in \mathcal{T}_{K,K'}$ and on the event $\Omega_{n,N,K_{\max}}$, one has $\|h\|_{n,N}^2 \leq \frac{3}{2}\|h\|_n^2 = \frac{3}{2}$. Then, using the chaining technique of [7], we have

$$\begin{aligned}
\widetilde{\mathbb{P}}\left(\sup_{h\in\mathcal{T}_{K,K'}} \nu_1(h) > \tau\right) &= \widetilde{\mathbb{P}}\left(\exists (h_k)_{k\geq 0} \in \prod_{k\geq 0} E_{\varepsilon_k} / \nu_1(h) = \nu_1(h_0) + \sum_{k=1}^{\infty}\nu_1(h_k - h_{k-1}) > \tau\right) \\
&\leq \sum_{h_0\in E_0} \widetilde{\mathbb{P}}\left(\nu_1(h_0) > y_0^{n,N}\right) + \sum_{k=1}^{\infty}\sum_{\substack{h_k\in E_{\varepsilon_k} \\ h_{k-1}\in E_{\varepsilon_{k-1}}}} \widetilde{\mathbb{P}}\left(\nu_1(h_k - h_{k-1}) > y_k^{n,N}\right).
\end{aligned}$$

According to Equation (2.71) and Lemma 2.8.10, there exists a constant $C > 0$ such that

$$\begin{aligned}
\widetilde{\mathbb{P}}\left(\nu_1(h_0) > y_0^{n,N}\right) &\leq \widetilde{\mathbb{P}}\left(\nu_1(h_0) > \sigma_1\sqrt{6x_0^{n,N}} + \sigma_1^2\|h_0\|_\infty x_0^{n,N}\right) \\
&\leq \exp(-CNnx_0^{n,N}), \\
\forall k \geq 1,\ \widetilde{\mathbb{P}}\left(\nu_1(h_k - h_{k-1}) > y_0^{n,N}\right) &\leq \widetilde{\mathbb{P}}\left(\nu_1(h_k - h_{k-1}) > \sigma_1\sqrt{6x_k^{n,N}} + \sigma_1^2\|h_k - h_{k-1}\|_\infty x_k^{n,N}\right) \\
&\leq \exp(-CNnx_k^{n,N}).
\end{aligned}$$

Finally, since $N_k = \log(|E_{\varepsilon_k}|)$ for all $k \geq 0$, we deduce that

$$\begin{aligned}
\widetilde{\mathbb{P}}\left(\sup_{h\in\mathcal{T}_{K,K'}} \nu_1(h) > \tau\right) &\leq |E_{\varepsilon_0}|\exp(-CNnx_0^{n,N}) + \sum_{k=1}^{\infty}\left(|E_{\varepsilon_k}| + \left|E_{\varepsilon_{k-1}}\right|\right)\exp(-CNnx_k^{n,N}) \\
&\leq \exp\left(N_0 - CNnx_0^{n,N}\right) + \sum_{k=1}^{\infty}\exp\left(N_k + N_{k-1} - CNnx_k^{n,N}\right).
\end{aligned}$$

We choose $x_0^{n,N}$ and $x_k^{n,N}$, $k \geq 1$ such that,

$$\begin{aligned}
N_0 - CNnx_0^{n,N} &= -a\left(K + K' + 2M\right) - b \\
N_k + N_{k-1} - CNnx_k^{n,N} &= -k(K + K' + 2M) - a\left(K + K' + 2M\right) - b
\end{aligned}$$

where a and b are two positive real numbers. We deduce that

$$x_k^{n,N} \leq C_0(1+k)\frac{K + K' + 2M}{Nn} \quad \text{and} \quad \tau \leq C_1\sigma_1^2\sqrt{\sqrt{L}\frac{K + K' + 2M}{Nn}} \tag{2.73}$$

with $C_0 > 0$ and C_1 two constants depending on a and b. It follows that

$$\widetilde{\mathbb{P}}\left(\sup_{t\in\mathcal{T}_{K,K'}} \nu(t) > \tau\right) \leq \frac{\mathrm{e}}{\mathrm{e}-1}\mathrm{e}^{-b}\exp\left\{-a\left(K + K' + 2M\right)\right\}.$$

From Equation (2.73), we set

$$q\left(K,K'\right)=\kappa^{*}\sigma_{1}^{2}\sqrt{L}\frac{K+K'+2M}{Nn}$$

where $\kappa^{*}>0$ depends on $C_{1}>0$. Thus, for all $K,K'\in\mathcal{K}$,

$$\mathbb{P}\left(\left\{\sup_{h\in\mathcal{T}_{K,K'}}\nu^{2}(h)>q\left(K,K'\right)\right\}\cap\Omega_{n,N,K_{\max}}\right)\leq\frac{\mathrm{e}^{-b+1}}{\mathrm{e}+1}\exp\left\{-a\left(K+K'+2M\right)\right\}$$

and there exists constants $c,C>0$ such that

$$\begin{aligned}\mathbb{E}\left[\left(G_{K}\left(K'\right)-q(K,K')\right)_{+}\mathbb{1}_{\Omega_{n,N,K_{\max}}}\right]&\\\leq\frac{c(K+K')}{Nn}\mathbb{P}\left(\left\{\sup_{t\in\mathcal{T}_{K,K'}}\nu^{2}(t)>q(K,K')\right\}\cap\Omega_{n,N,K_{\max}}\right)&\\\leq\frac{C}{Nn}\exp\left\{-\frac{a}{2}\left(K+K'\right)\right\}.&\end{aligned}$$

Finally, there exists a real constant $C>0$ such that,

$$\mathbb{E}\left[\left(G_{K}(\widehat{K})-q(K,\widehat{K})\right)_{+}\mathbb{1}_{\Omega_{n,N,K_{\max}}}\right]\leq\sum_{K'\in\mathcal{K}}\mathbb{E}\left[\left(G_{K}\left(K'\right)-q(K,K')\right)_{+}\mathbb{1}_{\Omega_{n,N,K_{\max}}}\right]\leq\frac{C}{Nn}.$$

We choose the penalty function pen such that for each $K\in\mathcal{K}$,

$$\mathrm{pen}(K)\geq\kappa\sigma_{1}^{2}\sqrt{L}\frac{K+M}{Nn}.$$

For N large enough, one has $\sigma_{1}^{2}\leq\sqrt{L}$. Thus, we finally set $\mathrm{pen}(K)=\kappa\frac{(K+M)\log(N)}{Nn}$ with $L=\log(N)$. Then, there exists a constant $C>0$ such that,

$$\mathbb{E}\left[\left\|\widehat{\sigma}^{2}_{\widehat{K}}-\sigma^{2}_{|I}\right\|^{2}_{n,N}\right]\leq 34\inf_{K\in\mathcal{K}}\left\{\inf_{h\in\mathcal{S}_{K+M}}\left\|h-\sigma^{2}_{|I}\right\|^{2}_{n}+\mathrm{pen}(K)\right\}+\frac{C}{Nn}.$$

□

Proof of Theorem 2.5.2

Proof. From Equation (2.23), we have

$$\widehat{K}:=\arg\min_{K\in\mathcal{K}}\gamma_{n,N}(\widehat{\sigma}^{2}_{K+M,L})+\mathrm{pen}(K).$$

Then, for all $K\in\mathcal{K}$ and for all $h\in\mathcal{S}_{K+M,L}$, we have

$$\gamma_{n,N}(\widehat{\sigma}^{2}_{\widehat{K},L})+\mathrm{pen}(\widehat{K})\leq\gamma_{n,N}(h)+\mathrm{pen}(K).$$

Then, for all $K\in\mathcal{K}$ and for all $h\in\mathcal{S}_{K+M,L}$,

$$\begin{aligned}\gamma_{n,N}(\widehat{\sigma}^{2}_{\widehat{K},L})-\gamma_{n,N}(\sigma^{2})&\leq\gamma_{n,N}(h)-\gamma_{n,N}(\sigma^{2})+\mathrm{pen}(K)-\mathrm{pen}(\widehat{K})\\\left\|\widehat{\sigma}^{2}_{\widehat{K},L}-\sigma^{2}\right\|^{2}_{n,N}&\leq\|h-\sigma^{2}\|^{2}_{n,N}+2\nu(\widehat{\sigma}^{2}_{\widehat{K},L}-h)+2\mu(\widehat{\sigma}^{2}_{\widehat{K},L}-h)+\mathrm{pen}(K)-\mathrm{pen}(\widehat{K}).\end{aligned}$$

We have for all $a>0$,

$$2\mathbb{E}\left[\mu\left(\widehat{\sigma}^{2}_{\widehat{K},L}-h\right)\right]\leq\frac{2}{a}\mathbb{E}\left\|\widehat{\sigma}^{2}_{\widehat{K},L}-\sigma^{2}\right\|^{2}_{n,N}+\frac{2}{a}\inf_{h\in\mathcal{S}_{K+M,L}}\|h-\sigma^{2}\|^{2}_{n}+\frac{a}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\mathbb{E}\left[\left(R^{j}_{k\Delta}\right)^{2}\right]$$

and since $\nu = \nu_1 + \nu_2 + \nu_3$, according to the proof of Theorem 2.4.3, there exists a constant $c > 0$ such that

$$\mathbb{E}\left[\nu(\widehat{\sigma}^2_{\widehat{K},L} - h)\right] \leq c\mathbb{E}\left[\nu_1(\widehat{\sigma}^2_{\widehat{K},L} - h)\right]$$

where the for $i \in \{1,2,3\}$ and for all $h \in \mathcal{S}_{K+M,L},\ K \in \mathcal{K}$,

$$\nu_i(h) = \frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1} h(X^j_{k\Delta})\zeta^j_{k\Delta},$$

and the $\zeta^j_{k\Delta}$ are given Then,

$$\left(1-\frac{2}{a}\right)\mathbb{E}\left[\left\|\widehat{\sigma}^2_{\widehat{K},L} - \sigma^2\right\|^2_{n,N}\right] \leq \left(1+\frac{2}{a}\right)\inf_{h\in\mathcal{S}_{K+M,L}}\left\|h-\sigma^2\right\|^2_n + 2c\mathbb{E}\left[\nu_1(\widehat{\sigma}^2_{\widehat{K},L} - h)\right]$$
$$+ \mathrm{pen}(K) - \mathrm{pen}(\widehat{K}) + \frac{a}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\mathbb{E}\left[\left(R^j_{k\Delta}\right)^2\right]$$

From Equation (2.37) and for $a = 4$, there exists a constant $C > 0$ such that

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{\widehat{K},L} - \sigma^2\right\|^2_{n,N}\right] \leq 3\inf_{h\in\mathcal{S}_{K+M,L}}\left\|h-\sigma^2\right\|^2_n + 4c\mathbb{E}\left[\nu_1(\widehat{\sigma}^2_{\widehat{K},L} - h)\right] + 2\left(\mathrm{pen}(K) - \mathrm{pen}(\widehat{K})\right) + C\Delta^2. \tag{2.74}$$

Since for all $K \in \mathcal{K},\ \mathrm{pen}(K) \geq 2\kappa^*\sigma_1^2(K+M)\sqrt{2L}/(Nn)$, define the function $q : (K,K') \mapsto q(K,K')$ such that

$$q(K,K') = 2C^*\sigma_1^2\frac{(K+K'+2M)\sqrt{2L}}{Nn} \geq 2\sigma_1^2 v\sqrt{x^{n,N}} + \sigma_1^2 v x^{n,N} \tag{2.75}$$

where

$$x^{n,N} \propto \left(\frac{K+K'+2M}{Nn}\right)^2 \ \text{ and } \ v = \sqrt{2L}.$$

The constant $C^* > 0$ depends on constants $\kappa^* > 0$ and $c > 0$ of Equation (2.74) such that

$$4cq(K,K') \leq \mathrm{pen}(K) + 2\mathrm{pen}(K').$$

Then for all $K \in \mathcal{K}$ and for all $h \in \mathcal{S}_{K+M,L}$,

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{\widehat{K},L} - \sigma^2\right\|^2_{n,N}\right] \leq 3\left(\inf_{h\in\mathcal{S}_{K+M,L}}\left\|h-\sigma^2\right\|^2_n + \mathrm{pen}(K)\right) + 4c\mathbb{E}\left[\left(\nu_1(\widehat{\sigma}^2_{\widehat{m},L} - h) - q(K,\widehat{K})\right)_+\right] + C\Delta^2. \tag{2.76}$$

For all $K \in \mathcal{K}$ and for all $h \in \mathcal{S}_{K+M,L}$ such that $\|h\|_\infty \leq \sqrt{L}$, we have ,

$$\left\|\widehat{\sigma}^2_{\widehat{K},L} - h\right\|^2_{n,N} \leq \left\|\widehat{\sigma}^2_{\widehat{K},L} - h\right\|^2_{\infty} \leq 2L =: v^2.$$

Then, using Equation (2.75) and Lemma 2.8.10, there exists a constant $C > 0$ such that for all $K, K' \in \mathcal{K}$ and for all $h \in \mathcal{S}_{K+M,L}$,

$$\mathbb{P}\left(\nu_1(\widehat{\sigma}^2_{K',L} - h) \geq q(K,K'),\ \|\widehat{\sigma}^2_{\widehat{K},L} - h\|^2_{n,N} \leq v^2\right) \leq \exp\left(-CNnx^{n,N}\right). \tag{2.77}$$

Since $L = \log(N)$, then for N large enough, $\sigma_1^2 \leq \sqrt{\log(N)}$, we finally choose

$$\mathrm{pen}(K) = \kappa\frac{(K+M)\log(N)}{Nn}$$

where $\kappa > 0$ is a new constant. Since $\mathbb{E}\left[\nu_1(\widehat{\sigma}^2_{\widehat{K},L} - h)\right] \leq \mathrm{O}\left(\sqrt{\frac{(K_{\max}+M)\log^2(N)}{Nn}}\right)$ (see proof of Theorem 2.4.3), for all $K \in \mathcal{K}$ and $h \in \mathcal{S}_{K+M,L}$, there exists a constant $c > 0$ such that

$$\mathbb{E}\left[\left(\nu_1(\widehat{\sigma}^2_{\widehat{K},L} - h) - q(K,\widehat{K})\right)_+\right] \leq \max_{K'\in\mathcal{K}}\left\{\mathbb{E}\left[\left(\nu_1(\widehat{\sigma}^2_{K',L} - h) - q(K,K')\right)_+\right]\right\}$$
$$\leq cq(K,K_{\max})\max_{K'\in\mathcal{K}}\left\{\mathbb{P}\left(\nu_1(\widehat{\sigma}^2_{K',L} - h) \geq q(K,K')\right)\right\}.$$

From Equation (2.77), we obtain that

$$\mathbb{E}\left[\left(\nu_1(\widehat{\sigma}^2_{\widehat{K},L} - h) - q(K,\widehat{K})\right)_+\right] \leq cq(K,K_{\max})\exp(-CNn) \leq \frac{C}{Nn} \tag{2.78}$$

since K and $K_{\max}$ increase with the size N of the sample paths $D_{N,n}$, and

$$cNnq(K,K_{\max})\exp(-CNn) \to 0 \text{ as } N \to \infty.$$

Then, from Equations (2.78) and (2.76), there exists a constant $C > 0$ such that

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{\widehat{K},L} - \sigma^2\right\|^2_{n,N}\right] \leq 3\inf_{K\in\mathcal{K}}\left\{\inf_{h\in\mathcal{S}_{K+M,L}}\left\|h-\sigma^2\right\|^2_n + \text{pen}(K)\right\} + \frac{C}{Nn}. \tag{2.79}$$

□

Proof of Theorem 2.5.3

Proof. The proof of Theorem 2.5.3 is similar to the proof of Theorem 2.5.1. Then, from Equation (2.70), for all $h \in \mathcal{S}_{K+M}$,

$$\left\|\widehat{\sigma}^2_{\widehat{K},L} - \sigma^2_{|I}\right\|^2_{n,1} \leq 3\left\|h-\sigma^2_{|I}\right\|^2_{n,1} + 20\sup_{h\in\mathcal{T}_{K,\widehat{K}}}\nu^2(h) + \frac{20}{n}\sum_{k=0}^{n-1}\left(R^1_{k\Delta}\right)^2 + 2\left(\text{pen}(K) - \text{pen}(\widehat{K})\right),$$

where $\mathcal{T}_{K,K'} = \{h \in \mathcal{S}_{K+M} + \mathcal{S}_{K'+M},\ \|h\|_X = 1,\ \|h\|_\infty \leq \sqrt{L}\}$. Let $q : \mathcal{K}^2 \longrightarrow \mathbb{R}_+$ such that $160q(K,K') \leq 18\text{pen}(K) + 16\text{pen}(K')$.

Recall that the $\mathbb{L}^2-$norm $\|.\|$, the norm $\mathbb{E}\left[\|.\|_X\right]$ and the empirical norm $\|.\|_n$ are equivalent on $\mathbb{L}^2(I)$ since the transition density is bounded on the compact interval I. Then, for all $K \in \mathcal{K}$ and $h \in \mathcal{S}_{K+M,L}$, we have

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{\widehat{K},L} - \sigma^2_{|I}\right\|^2_{n,1}\right] \leq 3\left(\inf_{h\in\mathcal{S}_{K+M,L}}\left\|h-\sigma^2_{|I}\right\|^2_n + \text{pen}(K)\right) + 20\mathbb{E}\left(\sup_{h\in\mathcal{T}_{K,\widehat{K}}}\nu^2_1(h) - q(K,\widehat{K})\right) + C\Delta^2$$

where

$$\nu_1(h) := \frac{1}{n}\sum_{k=0}^{n-1}h(X^1_{k\Delta})\zeta^{1,1}_{k\Delta}$$

with $\zeta^{1,1}_{k\Delta}$ the error term. We set for all $K, K' \in \mathcal{K}$, $G_K(K') := \sup_{h\in\mathcal{T}_{K,K'}}\nu^2_1(h)$. Then, there exists $C > 0$ such that

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{\widehat{K},L} - \sigma^2_{|I}\right\|^2_{n,1}\right] \leq 3\inf_{K\in\mathcal{K}}\left\{\inf_{h\in\mathcal{S}_{K+M,L}}\left\|h-\sigma^2_{|I}\right\|^2_n + \text{pen}(K)\right\} + 20\sum_{K'\in\mathcal{K}}\mathbb{E}\left[\left(G_K(K') - q(K,K')\right)_+\right] + C\Delta^2.$$

Considering the unit ball $\overline{B}_{\|.\|_X}(0,1)$ of the approximation subspace given by

$$\overline{B}_{\|.\|_X}(0,1) = \left\{h \in \mathcal{S}_{K+M},\ \|h\|^2_X \leq 1\right\} = \left\{h \in \mathcal{S}_{K+M},\ \|h\|^2 \leq \frac{1}{\tau_0}\right\}.$$

We obtain from the proof of Theorem 2.5.1 with $N = 1$ that,

$$\sum_{K'\in\mathcal{K}}\mathbb{E}\left[\left(G_K(K') - q(K,K')\right)_+\right] \leq \frac{C}{n},$$

where $C > 0$ is a constant, $q(K,K') \propto \sigma^4_1\frac{(K+K'+2M)\sqrt{\log(n)}}{n}$ and $\text{pen}(K) \propto \frac{(K+M)\log(n)}{n}$. Then we obtain

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2_{\widehat{K},L} - \sigma^2_{|I}\right\|^2_{n,1}\right] \leq \frac{3}{\tau_0}\inf_{K\in\mathcal{K}}\left\{\inf_{h\in\mathcal{S}_{K+M,L}}\left\|h-\sigma^2_{|I}\right\|^2_n + \text{pen}(K)\right\} + \frac{C}{n}.$$

□

2.9 Appendix

Calibration

Fix the drift function $b(x) = 1 - x$, the time-horizon $T = 1$ and at time $t = 0,\ x_0 = 0$. Consider the following three models:

Model 1: $\sigma(x) = 1$

Model 2: $\sigma(x) = 0.1 + 0.9/\sqrt{1 + x^2}$

Model 3: $\sigma(x) = 1/3 + \sin^2(2\pi x)/\pi + 1/(\pi + x^2)$.

The three diffusion models satisfy Assumption 2.2.1 and are used to calibrate the numerical constant κ of the penalty function given in Theorem 2.5.1

As we already know, the adaptive estimator of σ^2 on the interval $[-\sqrt{\log(N)}, \sqrt{\log(N)}]$ necessitate a data-driven selection of an optimal dimension through the minimization of the penalized least squares contrast given in Equation (2.23) . Since the penalty function $\mathrm{pen}(d_N) = \kappa(K_N + M)\log^2(N)/N^2$ depends on the unknown numerical constant $\kappa > 0$, the goal is to select an optimal value of κ in the set $\mathcal{V} = \{0.1, 0.5, 1, 2, 4, 5, 7, 10\}$ of its possible values. To this end, we repeat 100 times the following steps:

1. Simulate learning samples D_N and $D_{N'}$ with $N \in \{50, 100\}$, $N' = 100$ and $n \in \{100, 250\}$
2. For each $\kappa \in \mathcal{V}$:
   (a) For each $K_N \in \mathcal{K}$ and from D_N, compute $\widehat{\sigma}^2_{d_N, L_N}$ given in Equations (2.10) and (2.11).
   (b) Select the optimal dimension $\widehat{K}_N \in \mathcal{K}$ using Equation (2.23)
   (c) Using the learning sample $D_{N'}$, evaluate $\left\|\widehat{\sigma}^2_{\widehat{d}_N, L_N} - \sigma^2_A\right\|^2_{n,N'}$ where $\widehat{d}_N = \widehat{K}_N + M$.

Then, we calculate average values of $\left\|\widehat{\sigma}^2_{\widehat{d}_N, L_N} - \sigma^2_A\right\|^2_{n,N'}$ for each $\kappa \in \mathcal{V}$ and obtain the following results:

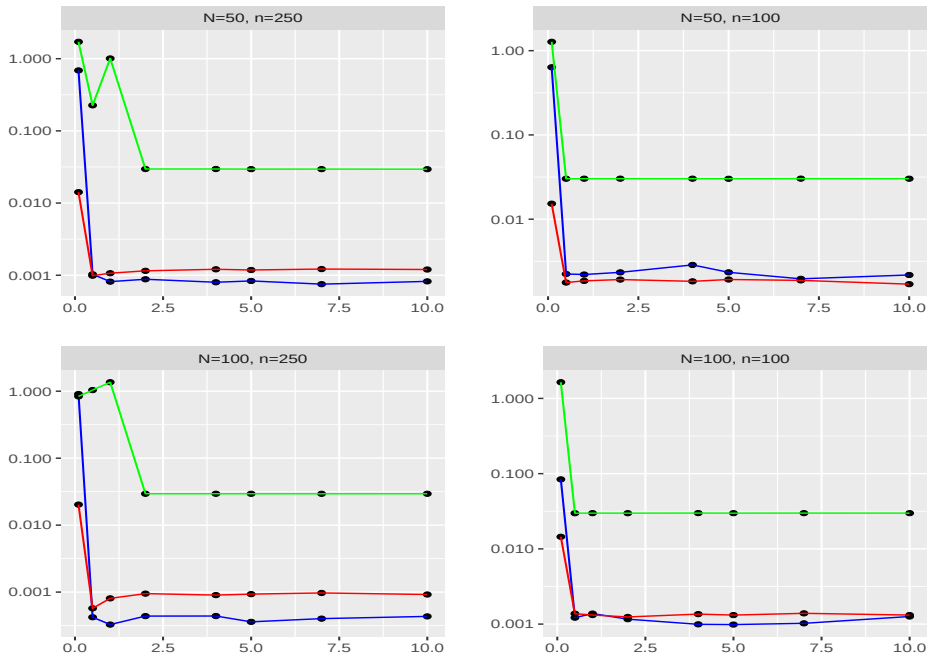

Figure 2.2: Calibration of the constant $\kappa \in \mathcal{V}$ of the penalty function

We finally choose $5 \in \mathcal{V}$ as the optimal value of κ in reference to the results of Figure 2.2 .

Proof of Lemma 2.8.10

Proof. We obtain from *Comte,Genon-Catalot,Rozenholc (2007) proof of Lemma 3* that for each $j \in [\![1, N]\!]$, $k \in [\![0, n-1]\!]$ and $p \in \mathbb{N} \setminus \{0, 1\}$

$$\mathbb{E}\left[\exp\left(ug(X_{k\Delta}^j)\xi_{k\Delta}^{j,1} - \frac{au^2g^2(X_{k\Delta}^j)}{1-bu}\right)|\mathcal{F}_{k\Delta}\right] \leq 1$$

with $a = \mathrm{e}\left(4\sigma_1^2c^2\right)^2$, $b = 4\sigma_1^2c^2\mathrm{e}\|g\|_\infty$, $u \in \mathbb{R}$ such that $bu < 1$ and $c > 0$ a real constant. Thus,

$$\begin{aligned}\mathbb{P}\left(\frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1} g(X_{k\Delta}^j)\xi_{k\Delta}^{j,1} \geq \varepsilon, \|g\|_{N,n}^2 \leq v^2\right) &= \mathbb{E}\left(\mathbf{1}_{\left\{\sum_{j=1}^{N}\sum_{k=0}^{n-1} ug(X_{k\Delta}^j)\xi_{k\Delta}^{(j,1)} \geq Nnu\varepsilon\right\}}\mathbf{1}_{\|g\|_{n,N}^2 \leq v^2}\right)\\ &= \mathbb{E}\left(\mathbf{1}_{\left\{\exp\left(\sum_{j=1}^{N}\sum_{k=0}^{n-1} ug(X_{k\Delta}^j)\xi_{k\Delta}^{(j,1)}\right)\mathrm{e}^{-Nnu\varepsilon} \geq 1\right\}}\mathbb{1}_{\|g\|_{N,n}^2 \leq v^2}\right)\\ &\leq \mathrm{e}^{-Nnu\varepsilon}\mathbb{E}\left[\mathbb{1}_{\|g\|_{n,N}^2 \leq v^2}\exp\left\{\sum_{j=1}^{N}\sum_{k=0}^{n-1} ug(X_{k\Delta}^j)\xi_{k\Delta}^{j,1}\right\}\right].\end{aligned}$$

It follows that,

$$\mathbb{P}\left(\frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1} g(X_{k\Delta}^j)\xi_{k\Delta}^{j,1} \geq \varepsilon, \|g\|_{n,N}^2 \leq v^2\right) \leq \quad \exp\left\{-Nnu\varepsilon + \tfrac{Nnau^2v^2}{1-bu}\right\}.$$

We set $u = \frac{\varepsilon}{\varepsilon b + 2av^2}$. Then, we have $-Nnu\varepsilon + Nnav^2u^2/(1-bu) = -Nn\varepsilon^2/2(\varepsilon b + 2av^2)$ and,

$$\begin{aligned}\mathbb{P}\left(\frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1} g(X_{k\Delta}^j)\xi_{k\Delta}^{j,1} \geq \varepsilon, \|g\|_{n,N}^2 \leq v^2\right) &\leq \exp\left(-\frac{Nn\varepsilon^2}{2(\varepsilon b + av^2)}\right)\\ &\leq \exp\left(-C\frac{Nn\varepsilon^2}{\sigma_1^2\left(\varepsilon\|g\|_\infty + 4\sigma_1^2v^2\right)}\right)\end{aligned}$$

where $C > 0$ is a constant depending on $c > 0$. □

Proof of Lemma 2.8.9

Proof. Set $K_{n,N} = K_N$ since $N \propto n$. Let us remind the reader of the Gram matrix Ψ_{K_N} given in Equation (2.20),

$$\Psi_{K_N} = \mathbb{E}\left[\frac{1}{Nn}\mathbf{F}'_{K_N}\mathbf{F}_{K_N}\right] = \mathbb{E}\left(\widehat{\Psi}_{K_N}\right)$$

where,

$$\mathbf{F}_{K_N} := \left((B_\ell(X_0^j), \ldots, (B_\ell(X_{(n-1)\Delta}^j))\right)_{\substack{0 \leq \ell \leq K_N-1\\ 1 \leq j \leq N}} \in \mathbb{R}^{Nn \times (K_N+M)} \tag{2.80}$$

The empirical counterpart $\widehat{\Psi}$ is the random matrix given by $\widehat{\Psi}_{K_N}$ of size $(K_N + M) \times (K_N + M)$ is given by

$$\widehat{\Psi}_{K_N} := \frac{1}{Nn}\mathbf{F}'_{K_N}\mathbf{F}_{K_N} = \left(\frac{1}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1} f_\ell(X_{k\Delta}^j)f_{\ell'}(X_{k\Delta}^j)\right)_{\ell,\ell' \in [-M, K_N-1]}. \tag{2.81}$$

For all $t = \sum_{\ell=-M}^{K_N-1} a_\ell B_{\ell,M,\mathbf{u}} \in S_{K_N,M}$ one has

$$\|t\|_{n,N}^2 = a'\widehat{\Psi}_{K_N}a \text{ and } \|t\|_n^2 = a'\Psi_{K_N}a, \text{ with } a = (a_{-M}, \cdots, a_{K_N-1})'.$$

Under Assumption 2.2.1, we follow the lines of [21] *Proposition 2.3* and *Lemma 6.2*. Then,

$$\begin{aligned}\sup_{t\in S_{K_N,M},\|t\|_n=1}\left|\|t\|_{n,N}^2-\|t\|_n^2\right| &= \sup_{w\in\mathbb{R}^{K_N+M},\left\|\Phi_{K_N}^{1/2}w\right\|_{2,K_N+M}=1}\left|w'\left(\widehat{\Psi}_{K_N}-\Psi_{K_N}\right)w\right| \\ &= \sup_{u\in\mathbb{R}^{K_N+M},\|u\|_{2,K_N+M}=1}\left|u'\Psi_{K_N}^{-1/2}\left(\widehat{\Psi}_{K_N}-\Psi_{K_N}\right)\Psi_{K_N}^{-1/2}u\right| \\ &= \left\|\Psi_{K_N}^{-1/2}\widehat{\Psi}_{K_N}\Psi_{K_N}^{-1/2}-\mathrm{Id}_{K_N+M}\right\|_{\mathrm{op}}.\end{aligned}$$

Therefore,

$$\Omega_{n,N,K_N}^c=\left\{\left\|\Psi_{K_N}^{-1/2}\widehat{\Psi}_{K_N}\Psi_{K_N}^{-1/2}-\mathrm{Id}_{K_N+M}\right\|_{\mathrm{op}}>1/2\right\}.$$

Since $A_N=o\left(\sqrt{\log(N)}\right)$, we obtain from [28], *proof of Lemma 7.8*, there exists a constant $C>0$ such that

$$\mathbb{P}\left(\Omega_{n,N,K_N}^c\right)\leq 2(K_N+M)\exp\left(-C\log^{3/2}(N)\right). \tag{2.82}$$

Finally, since $2(K_N+M)\exp\left(-(C/2)\log^{3/2}(N)\right)\longrightarrow 0$ as $N\longrightarrow+\infty$, one concludes from Equation (2.82) and for N large enough,

$$\mathbb{P}\left(\Omega_{n,N,K_N}^c\right)\leq C\exp\left(-c\log^{3/2}(N)\right)$$

where $c>0$ and $C>0$ are new constants. □

Chapter 3

Procédure non-paramétrique de classification multiclasses de type *plug-in* pour les trajectoires de diffusions

Résumé Nous étudions le problème de classification multiclasses où la variable vient d'un mélange de diffusions homogènes en temps. Spécifiquement, les classes sont discriminées par leurs fonctions de dérive tandis que le coefficient de diffusion est commun à toutes les classes et inconnu. Dans ce cadre, nous construisons un classifieur de type *plug-in* qui repose sur des estimateurs non-paramétriques des fonctions de dérive et de diffusion. Nous établissons premièrement la consistance de notre procédure de classification sous des hypothèses faibles sur le modèle étudié, et ensuite, nous fournissons des vitesses de convergence sous des ensembles d'hypothèses différents. Finalement, une étude numérique soutient nos résultats théoriques.

Mots clés: Apprentissage supervisé; Classification multiclasses; Estimation non-paramétrique; Classifieur de type *plug-in*; Processus de diffusion

Abstract : We study the multiclass classification problem where the features come from a mixture of time-homogeneous diffusions. Specifically, the classes are discriminated by their drift functions while the diffusion coefficient is common to all classes and unknown. In this framework, we build a plug-in classifier which relies on nonparamateric estimators of the drift and diffusion functions. We first establish the consistency of our classification procedure under mild assumptions and then provide rates of convergence under different set of assumptions. Finally, a numerical study supports our theoretical findings.

Keywords: Supervised learning; Multiclass classification; Nonparametric estimation; Plug-in classifier; Diffusion process

Contents

3.1 Introduction

The massive collection of functional data has found many applications in recent years for the modeling of the joint (time)-evolution of agents – individuals, species, particles – that are represented by some sets of features – time-varying variables such as geographical positions, population sizes, portfolio values etc. Examples can be found in mathematical finance (see e.g. [34]), biology (see e.g. [36]), or physics (see e.g. [33]). This gave rise to an abundant literature on statistical methods for functional data, (see e.g. [79, 93], for a review). Within this context, the study of efficient supervised classification procedures that are designed to handle temporal data is a major challenge. Indeed, usual learning algorithms such as random forests, kernel methods or neural networks are not directly tailored to take into account the temporal dependency of the data. Recently, this question has drawn a lot of attention, see [24, 5, 94, 26, 64] any references therein.

In the present paper, we tackle the multiclass classification problem where the features belong to a particular family of functional data, namely trajectories, whose temporal dynamic is modelled by stochastic differential equation. In this framework, we propose a nonparametric plug-in type procedure for such data generated by diffusion processes observed at discrete time. Hence, our work takes place in the high frequency setup. Let us denote by (X, Y) a random couple built on a probability space $\left(\Omega, \mathcal{F}, \mathbb{P}_{(X,Y)}\right)$. The feature $X = (X_t)_{t\in[0,1]}$ is a real-valued diffusion process whose drift coefficient depends on its associated label Y taking values in $\mathcal{Y} = \{1, \cdots, K\}$, with $K \geq 2$. More precisely, for each $i \in \mathcal{Y}$, X is a solution of a stochastic differential equation whose drift function, denoted by b_i^*, depends on the class i. The marginal distribution of X is hence a mixture of distributions of time-homogeneous diffusion processes. We assume that a learning sample $\mathcal{D}_N = \{((X_t^i)_{t\in[0,1]}, Y_i), i = 1, \ldots, N\}$ is provided, composed of N *i.i.d.* random couples with distribution $\mathbb{P}_{(X,Y)}$. Additionally, in this paper, the diffusions X_i are observed on a subdivision $\{0, 1/n, \cdots, 1\}$ of the time interval $[0, 1]$, for a positive integer n. Since we deal with multiclass classification setting, the statistical goal is then to build, based on $\mathcal{D}_N$, a classifier $\widehat{g}$, such that $\widehat{g}(X)$ is a prediction of the associated label Y of a new path X. Besides, we expect that the empirical classifier mimics the optimal Bayes classifier g^* characterized as

$$g^*(X) \in \arg\min_g \mathbb{P}_{(X,Y)}\left(g(X) \neq Y\right).$$

Specifically, we propose a classification procedure based on the plug-in principle. In particular, the construction of our empirical classifier relies on estimators of both drift and diffusion coefficients. The performance of a predictor $\widehat{g}$ is assessed through its excess risk $\mathbb{P}(\widehat{g}(X) \neq Y) - \mathbb{P}\left(g^*(X) \neq Y\right)$. In the finite dimensional classification setup (*e.g.* $X \in \mathbb{R}^d$), rates of convergence for plug-in rules are

usually obtained under the strong density assumption (X admits a density which is lower bounded) as in [4, 45]. However, theoretical properties of plug-in rules in supervised classification of trajectories are much less studied.

Related works. Up to our knowledge, the work of [11] is the first one that tackles the problem of supervised classification in the stochastic differential equation framework. More precisely, the authors consider the model where $X = (X_t)_{t\in[0,1]}$ is a mixture of two diffusion processes and provide a classifier based on the empirical risk minimization strategy for which they establish rates of convergence. However, the proposed method is not implementable since it involves the minimization of a non-convex criterion. More recently, [44], and [29] study plug-in classifiers for classification of diffusion paths. In [44] the authors propose a plug-in rule for the binary classification problem where the trajectories are generated by Gaussian processes, solutions of the white noise model. In this model, the drift function depends on time and on the label Y, also, the diffusion coefficient is supposed to be constant and known. Within this framework, [44] establish the optimality of their classification procedure which reaches the minimax rate of convergence of order $N^{-s/(2s+1)}$, where the drift function is assumed to belong to a Sobolev space of regularity $s \geq 1$. Under an additional margin type assumption, they also derive faster rates of convergence. Closest to our framework, [29] also consider the challenging multiclass problem where the drift functions are space-dependent. However, the authors consider drift functions modeled under parametric assumptions, keeping the diffusion coefficient known and constant. They propose a plug-in classifier for which only consistency is established.

In the present work, we consider a plug-in classifier that relies on nonparametric estimators of the drift and diffusion coefficients. The literature on this topic is extensive. Usually, the construction of estimators of drift and diffusion functions relies on the observation of a single path. For instance, [58] studies minimax rate of convergence for the estimation of the diffusion coefficient on a compact interval. For the inference of the drift coefficient, the main references using penalized contrasts can be found for long time observation with high frequency data in [57, 22, 20]. However, since we deal with the multiclass classification framework, the construction of estimators of both drift and diffusion coefficients is based on the learning sample $\mathcal{D}_N$ which is composed of repeated observations of the process on the fixed time-interval $[0, 1]$. Recently, [18, 70, 27] consider nonparametric procedures for the estimation of the drift function for continuous observations in the context of *i.i.d.* observations when the horizon time is fixed. Furthermore, towards high-frequency data, [30] study minimum contrast estimator under a l_2 constraint.

Main contributions. In this paper, we extend the results of [29] and [44] in several directions. In particular, one of the major contribution is to provide, up to our knowledge, the first study of rates of convergence for plug-in classifier in the mixture model of time-homogeneous diffusion. Importantly, we highlight that extending the results of [44] to diffusion models in which the drift functions are space-dependent and the diffusion coefficient is either unknown or non-constant add many difficulties. Besides, contrary to [29], we consider the nonparametric mixture model where both drift *and* diffusion functions are unknown as well as the weights of the mixture. Specifically, we build a plug-in classifier that relies on the Girsanov's theorem and involves nonparametric estimators of the drift functions $b_i^*, i \in \mathcal{Y}$, and the diffusion coefficient. The construction of our estimators is inspired of the ridge estimators provided in [30], and consists in the minimization of a least-squares type contrast over a finite dimensional subspace under a l_2-constraint. The considered space of approximation is then spanned by the B-spline basis [25].

One of the main difficulty of the study of statistical properties of the plug-in classifiers in our context is that it requires deriving rates of convergence for the drift and diffusion coefficients on a non-compact interval. It hence implies that the strong density assumption does not hold, although, we consider assumptions that ensure existence of transition density. Notably, our results embed generalization of the results provided in [30] for the estimation of non-compactly supported drift functions for B-spline based estimators, but also exhibit the first result for the estimation of the diffusion coefficient in the *i.i.d.* framework. A salient point of our theoretical findings is obtained when the diffusion

coefficient is constant and known. In this case, by leveraging the results of [18], we show that optimal rates for drift estimation can only be achieved on intervals included in $[-C\sqrt{\log(N)}, C\sqrt{\log(N)}]$, with $C > 0$.

To sum up our results, a first part is dedicated to the consistency of our plug-in classifier which is obtained under very mild assumptions. In a second part, convergence rates are established in three particular cases.

(i) When the drift functions are bounded and Lipschitz, and the diffusion coefficient is unknown and possibly non-constant, we obtain a rate of convergence of order $N^{-1/5}$ for the plug-in classifier (up to a factor of order $\exp(\sqrt{c\log(N)}),\ c > 0$).

(ii) When the diffusion coefficient is known and constant, and when the drift functions are bounded and belongs to some Hölder space with regularity β, using some arguments developed in [19] and [20] for the estimation of non-compactly supported drift functions, together with approximations of the transition density of X (as they are intractable), we then prove that the plug-in classifier reaches rate of order $N^{-\beta/(2\beta+1)}$ (up to a factor of order $\exp(\sqrt{c\log(N)}),\ c > 0$).

(iii) When the drifts are unbounded but re-entrant and Hölder continuous with regularity β, we obtain a rate of convergence of order $N^{-3\beta/(4(2\beta+1))}$. Notice that when $\beta = 1$ and $d = 1$, it corresponds to the rate found in [45].

3.2 Statistical setting

We consider the multiclass classification problem, where the feature X comes from a mixture of Brownian diffusions with drift. More precisely, the generic data-structure is a couple (X, Y) where the label Y takes its values in the set $\mathcal{Y} := \{1, \dots, K\}$ with distribution denoted by $\mathfrak{p}^* = (\mathfrak{p}_1^*, \cdots, \mathfrak{p}_K^*)$, and where the process $X = (X_t)_{t\in[0,1]}$ is defined as the solution of the following stochastic differential equation

$$dX_t = b_Y^*(X_t)dt + \sigma^*(X_t)dW_t, \quad X_0 = 0, \tag{3.1}$$

where $(W_t)_{t\geq 0}$ is a standard Brownian motion independent of Y. In the following, we denote by $\mathbf{b}^* = (b_1^*, \dots, b_K^*)$ the vector of drift functions. The real-valued functions $b_i^*(.)$, $i \in \mathcal{Y}$, and the diffusion coefficient $\sigma^*(.)$ are assumed to be unknown. We also assume that $0 < \mathfrak{p}_0^* = \min_{i\in\mathcal{Y}} \mathfrak{p}_i^*$.

In this framework, the objective is to build a classifier g, *i.e.* a measurable function such that the value $g(X)$ is a prediction of the associated label Y of X. The accuracy of such classifier g is then assessed through its misclassification risk, denoted by

$$\mathcal{R}(g) := \mathbb{P}_{(X,Y)}\left(g(X) \neq Y\right).$$

In the following, the set of all classifiers is denoted by $\mathcal{G}$.

The main assumptions considered throughout the paper are presented in Section 3.2.1. The definition and characterization of the optimal classifier *w.r.t.* the misclassification risk, namely the *Bayes classifier*, is provided in Section 3.2.2

3.2.1 Assumptions

The following assumptions ensure that Equation (3.1) admits a unique strong solution (see [63], Theorem 2.9), and that the diffusion process X admits a transition density

$$p_X : (t, x) \in ([0, 1] \times \mathbb{R}) \mapsto p_X(t, x)$$

(see for example [52]).

Assumption 3.2.1. *(Ellipticity and regularity)*

(i) There exists $L_0 > 0$ such that the functions $b_i^*, i = 1, \dots, K$ and σ^* are L_0-Lipschitz:

$$\sup_{i\in\mathcal{Y}} |b_i^*(x) - b_i^*(y)| + |\sigma^*(x) - \sigma^*(y)| \leq L_0|x - y|,\ \forall (x, y) \in \mathbb{R}^2.$$

(ii) There exist real constants σ_0^*, σ_1^* *such that*

$$0 < \sigma_0^* \leq \sigma^*(x) \leq \sigma_1^*, \quad \forall x \in \mathbb{R}.$$

(iii) $\sigma^* \in \mathcal{C}^2(\mathbb{R})$ *and there exist* $\gamma \geq 0$ *and* $c > 0$ *such that :* $|\sigma^{*\prime}(x)| + |\sigma^{*\prime\prime}(x)| \leq c\left(1 + |x|^\gamma\right), \quad \forall x \in \mathbb{R}.$

Assumption 3.2.1 insures that for any integer $q \geq 1$, there exists $C_q > 0$ such that

$$\mathbb{E}_X\left[\sup_{t\in[0,1]} |X_t|^q\right] \leq C_q.$$

We also assume that the following Novikov's criterion is fulfilled [81, Prop. (1.15) p. 308] .

Assumption 3.2.2. *(Novikov's condition) For all* $i \in \mathcal{Y}$*, we have*

$$\mathbb{E}_X\left[\exp\left(\frac{1}{2}\int_0^1 \frac{b_i^{*2}}{\sigma^{*2}}(X_s)ds\right)\right] < +\infty.$$

In particular, this assumption allows to apply Girsanov's theorem that is a key ingredient to derive a characterization of the Bayes classifier in the next section.

3.2.2 Bayes Classifier

The Bayes classifier g^* is a minimizer of the misclassification risk over $\mathcal{G}$

$$g^* \in \arg\min_{g\in\mathcal{G}} \mathcal{R}(g),$$

and is expressed as

$$g^*(X) \in \operatorname*{argmax}_{i\in\mathcal{Y}} \pi_i^*(X), \quad \text{with} \quad \pi_i^*(X) := \mathbb{P}_{(X,Y)}\left(Y = i|X\right).$$

The following result of [29] provides a closed form of the conditional probabilities π_i^*, $i \in \mathcal{Y}$.

Proposition 3.2.3. *[29] Under Assumptions 3.2.1, 3.2.2, for all* $i \in \mathcal{Y}$*, we define*

$$F_i^*(X) := \int_0^1 \frac{b_i^*}{\sigma^{*2}}(X_s)dX_s - \frac{1}{2}\int_0^1 \frac{b_i^{*2}}{\sigma^{*2}}(X_s)ds.$$

Under Assumptions 3.2.1,3.2.2, for each $i \in \mathcal{Y}$*, the conditional probability* π_i^* *is given as follows:*

$$\pi_i^*(X) = \phi_i^*\left(\mathbf{F}^*(X)\right),$$

where $\mathbf{F}^* = (F_1^*, \ldots, F_k^*)$*, and* $\phi_i^* : (x_1, \cdots, x_K) \mapsto \frac{\mathfrak{p}_i^* e^{x_i}}{\sum_{k=1}^K \mathfrak{p}_k^* e^{x_k}}$ *are the softmax functions.*

The above proposition provides an explicit dependency of the Bayes classifier on the unknown parameters $\mathbf{b}^*$, σ^*, and $\mathfrak{p}^*$. Hence, it naturally suggests to build *plug-in* type estimators $\widehat{g}$ of the Bayes classifier g^*, relying on estimators of the unknown parameters. In this way, we aim at building an empirical classifier whose misclassification risk is closed to the minimum risk which is reached by the Bayes classifier. The following section is devoted to the presentation of the classification procedure.

3.3 Classification procedure: a plug-in approach

Let $n \geq 1$ be an integer, and $\Delta_n = 1/n$ the time step which defines the regular grid of the observation time interval $[0, 1]$. Let us assume now that an observation is a couple $(\bar{X}, Y)$, with $\bar{X} := (X_{k\Delta_n})_{0\leq k\leq n}$ a high frequency sample path coming from $(X_t)_{t\in[0,1]}$ a solution of Equation (3.1), and Y its associated label. We also introduce, for $N \geq 1$, a learning dataset $\mathcal{D}_N = \{(\bar{X}^j, Y_j),\ j \in \{1, \ldots N\}\}$ which consists of N independent copies of $(\bar{X}, Y)$. The asymptotic framework is such that N and n tend to infinity.

Based on $\mathcal{D}_N$ we build a classification procedure that relies on the result of Proposition 3.2.3. Our classifier uses the knowledge of the class Y_j for the path X^j, placing our work in the frame of supervised learning. The procedure is formally described in Section 3.3.1 and Section 3.3.2 while its statistical properties are provided in Section 3.3.3.

3.3.1 Classifier and excess risk

As suggested by Proposition 3.2.3, based on $\mathcal{D}_N$, we first build estimators $\widehat{\boldsymbol{b}} = (\widehat{b}_1, \dots, \widehat{b}_K)$, and $\widehat{\sigma}^2$ of $\mathbf{b}^*$ and σ^{*2} respectively. Besides, we consider the empirical estimators of $\mathfrak{p}_i^*$, $i = 1, \dots, K$:

$$\widehat{\mathfrak{p}}_i = \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}_{\{Y_j = i\}}. \tag{3.2}$$

Then, in a second step, we introduce the discretized estimator of $\mathbf{F}^*$

$$\widehat{\boldsymbol{F}} = (\widehat{F}_1, \dots, \widehat{F}_K), \quad \text{with} \quad \widehat{F}_i(X) = \sum_{k=0}^{n-1} \left(\frac{\widehat{b}_i}{\widehat{\sigma}^2}(X_{k\Delta}) \left(X_{(k+1)\Delta} - X_{k\Delta} \right) - \frac{\Delta}{2} \frac{\widehat{b}_i^2}{\widehat{\sigma}^2}(X_{k\Delta}) \right). \tag{3.3}$$

Finally, considering the functions $\widehat{\phi}_i : (x_1, \cdots, x_K) \mapsto \frac{\widehat{\mathfrak{p}}_i e^{x_i}}{\sum_{k=1}^K \widehat{\mathfrak{p}}_k e^{x_k}}$, we naturally define the resulting plug-in classifier $\widehat{g}$ as

$$\widehat{g}(X) \in \underset{i \in \mathcal{Y}}{\operatorname{argmax}}\, \widehat{\pi}_i(X), \quad \text{with} \quad \widehat{\pi}_i(X) = \widehat{\phi}_i(\widehat{\boldsymbol{F}}(X)). \tag{3.4}$$

Hereafter, we establish that the consistency of the plug-in classifier $\widehat{g}$ can be obtained through an empirical distance between estimators $\widehat{\boldsymbol{b}}$, and $\widehat{\sigma}^2$ and the true functions $\mathbf{b}^*$, and σ^{*2} respectively. This distance relies on the empirical norm defined for $h : \mathbb{R} \to \mathbb{R}$ as

$$\|h\|_{n,i}^2 := \mathbb{E}_{X|Y=i} \left[\frac{1}{n} \sum_{k=0}^{n-1} h^2(X_{k\Delta}) \right].$$

We also introduce the general empirical norm $\|.\|_n$ which, for any function h, is

$$\|h\|_n^2 := \mathbb{E}_X \left[\frac{1}{n} \sum_{k=0}^{n-1} h^2(X_{k\Delta}) \right].$$

Let us begin with a proposition which establishes a closed formula of the excess risk in multiclass classification.

Proposition 3.3.1 ([29]). *Let g be a classifier. The following holds*

$$\mathcal{R}(g) - \mathcal{R}(g^*) = \mathbb{E} \left[\sum_{i=1}^{K} \sum_{j \neq i} \left| \pi_i^*(X) - \pi_j^*(X) \right| \mathbb{1}_{\{g(X)=j, g^*(X)=i\}} \right].$$

The proof of this result is omitted and can be found for instance in [29]. From the result of Proposition 3.3.1, and upper-bounding the indicator function by 1, we take advantage of the Lipschitz property of the softmax functions $(\phi_i^*)_{i=1,\dots,K}$ that define the probabilities $(\pi^*(X))_{i=1,\dots,K}$ to majorate the excess risk of an empirical classifier $\widehat{g}$ based on $\widehat{\mathbf{b}} = \left(\widehat{b}_1, \dots, \widehat{b}_K\right)$ and $\widehat{\sigma}^2$ by the respectives risks of estimation of estimators $\widehat{b}_i$ and $\widehat{\sigma}^2$. Let us now announce the main result on the excess risk of a plug-in type classifier.

Theorem 3.3.2. *Assume N and n fixed (and large). Grant Assumptions 3.2.1, 3.2.2. Assume that there exists $b_{\max}, \sigma_0^2 > 0$ such that for all $x \in \mathbb{R}$*

$$\max_{i \in \mathcal{Y}} |\widehat{b}_i(x)| \leq b_{\max} \quad \text{and} \quad \widehat{\sigma}^2(x) \geq \sigma_0^2. \tag{3.5}$$

Then the classifier $\widehat{g}$ defined in Equation (3.4) satisfies

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C \left(\sqrt{\Delta_n} + \frac{1}{\mathfrak{p}_0^* \sqrt{N}} + \mathbb{E}\left[b_{\max} \sigma_0^{-2} \sum_{i=1}^{K} \|\widehat{b}_i - b_i^*\|_n \right] + \mathbb{E}\left[\sigma_0^{-2} \|\widehat{\sigma}^2 - \sigma^{*2}\|_n \right] \right),$$

where $C > 0$ is a constant which depends on b^*, σ^*, and K.

Theorem 3.3.2 highlights that the excess risk of the plug-in classifier depends on the discretization error which is of order $\Delta_n^{-1/2}$, the L_2 error of $\widehat{\mathfrak{p}}$ which is of order $N^{-1/2}$, and the estimation error of $\widehat{\mathbf{b}}$ and $\widehat{\sigma}^2$ assessed through the empirical norm $\|.\|_n$. Therefore, a straightforward consequence of Theorem 3.3.2 is that consistent estimators of $\mathbf{b}^*$, and σ^{*2} yield the consistency of plug-in classifier $\widehat{g}$. Notice that the additional assumption (3.5) does not require that the true functions b_i^*'s are bounded, only their estimators should be. For the difference between b_i^* and $\widehat{b}_i$ to remain controlled in the norm $\|\cdot\|_n$, it is necessary that the process X rests with high probability in a compact region of $\mathbb{R}$. The next section is devoted to the construction of consistent estimators of both drift and diffusion coefficients.

3.3.2 Estimators of drift and diffusion coefficients

In this section, we provide consistent estimators $\widehat{\boldsymbol{b}}$, and $\widehat{\sigma}^2$, implying the consistency of the associated plug-in classifier. These estimators are defined as minimum contrast estimators under an l_2-constraint on a finite dimensional vector space spanned by the B-spline basis, but other families of nonparametric estimators could have been chosen as well. In particular, to ensure statistical guarantees on $\mathbb{R}$, the considered estimators are built on a large intervals parameterized by the number N of sample paths, and that tends to the whole real line as N goes to infinity.

Spaces of approximation

Let $A, \Xi > 0$, and $M \geq 1$. Let $\mathbf{u} = (u_{-M}, \ldots, u_{\Xi+M})$, a sequence of knots of the compact interval $[-A, A]$ such that

$$u_{-M} = \cdots = u_{-1} = u_0 = -A, \text{ and } u_\Xi = u_{\Xi+1} = \cdots = u_{\Xi+M} = A.$$
$$\forall \ell \in [\![0, \Xi]\!], \quad u_\ell = -A + \frac{2\ell A}{\Xi}.$$

Let us consider the B-spline basis $(B_{-M}, \ldots, B_{\Xi+M})$ of order M defined by the knots sequence $\mathbf{u}$. For the construction of the B-spline and its properties, we refer for instance to [54]. Let us mention that the considered B-spline functions are nonnegative and M-1 continuously differentiable on $[-A, A]$ and are zero outside $[-A, A]$. Besides, for all $x \in [-A, A]$, we have that $\sum_{\ell=-M}^{\Xi-1} B_\ell(x) = 1$. Now, we introduce the space of approximation $\mathcal{S}_{\Xi,M}$ defined as

$$\mathcal{S}_{\Xi,M} := \left\{ \sum_{\ell=-M}^{\Xi-1} a_\ell B_\ell, \quad \|\mathbf{a}\|_2^2 \leq (\Xi + M) A^2 \log(N) \right\}, \tag{3.6}$$

where $\|\mathbf{a}\|_2^2 = \sum_{\ell=-M}^{\Xi-1} a_\ell^2$ is the usual ℓ_2-norm. Note that A can depend on the size N of the learning sample and tend to infinity as $N \to \infty$. The introduction of the constraint space $\mathcal{S}_{\Xi,M}$ is motivated by two facts. The first one is the following important property of spline approximations, inspired by the related properties for the Hölder functions (see [54]):

Proposition 3.3.3. *Let h be a L-lipschitz function. Then there exists $\tilde{h} \in \mathcal{S}_{\Xi,M}$, such that*

$$|\tilde{h}(x) - h(x)| \leq C \frac{A}{\Xi}, \quad \forall x \in [-A, A],$$

where $C > 0$ depends on L, and M.

The second one is that the set of functions $\mathcal{S}_{\Xi,M}$ is a *totally bounded class*, in the following sense [32, Chapter 28]. According to [30], for each $\varepsilon \in (0, 1)$ and for N large enough, there exists an ε-net $\tilde{\mathcal{S}}_\varepsilon$ of $\mathcal{S}_{\Xi,M}$ *w.r.t.* to the supremum norm $\|.\|_\infty$ such that

$$\log\left(\text{card}(\tilde{\mathcal{S}}_\varepsilon)\right) \leq C_M \Xi \log\left(\frac{\Xi}{\varepsilon}\right).$$

It shows that the complexity of $\mathcal{S}_{\Xi,M}$ given in Equation (3.6) is parametric which is particularly appealing in order to apply concentration inequalities.

Minimum contrast estimators

In this section, we propose two estimators of $\boldsymbol{b}^*$, and σ^{*2} which lead to a plug-in classifier that exhibits appealing properties. The construction of the estimators $\widehat{\boldsymbol{b}}$, and $\widehat{\sigma}^2$ relies on the minimization of a least squares contrast function over the space $\mathcal{S}_{\Xi,M}$. They are both based on the observed increments of the process X.

Estimator of the drift functions. Let $i \in \mathcal{Y}$ and $N_i := \sum_{j=1}^{N} \mathbb{1}_{\{Y_j=i\}}$ a random variable of Binomial distribution with parameters $(N, \mathfrak{p}_i^*)$. We define the random set $\mathcal{I}_i := \{j,\ Y_j = i\} = \{i_1, \dots, i_{N_i}\}$ and consider the dataset $\left\{\bar{X}^j,\ j \in \mathcal{I}_i\right\}$ of size N_i composed of the observations of the class i. Hereafter, we work conditional on $(\mathbb{1}_{\{Y_1=i\}}, \dots, \mathbb{1}_{\{Y_N=i\}})$, on the event $\{N_i > 1\}$. Hence, N_i is viewed as a deterministic variable such that $N_i > 1$. In this context, we set for all $i \in \mathcal{Y}$, $A = A_{N_i} > 0$ and $\Xi = K_{N_i} > 0$ where (A_{N_i}) and (K_{N_i}) are increasing sequences of N_i. The first estimator $\widetilde{b}_i$ of b_i^* is defined as

$$\tilde{b}_i \in \underset{h \in \mathcal{S}_{K_{N_i},M}}{\arg\min} \frac{1}{nN_i} \sum_{j \in \mathcal{I}_i} \sum_{k=0}^{n-1} \left(Z^j_{k\Delta_n} - h(X^j_{k\Delta_n})\right)^2 \mathbb{1}_{N_i>0}, \quad \text{with} \quad Z^j_{k\Delta_n} := \frac{(X^j_{(k+1)\Delta_n} - X^j_{k\Delta_n})}{\Delta_n}. \tag{3.7}$$

Then, to fit the assumption of Theorem 3.3.2, rather than $\tilde{b}_i$, we consider its thresholded counterpart

$$\widehat{b}_i(x) := \tilde{b}_i(x)\mathbb{1}_{\{|\tilde{b}_i(x)| \le A_{N_i} \log^{1/2}(N)\}} + \operatorname{sgn}(\tilde{b}_i(x)) A_{N_i} \log^{1/2}(N) \mathbb{1}_{\{|\tilde{b}_i(x)| > A_{N_i} \log^{1/2}(N)\}}. \tag{3.8}$$

Note that the value of the threshold $A_{N_i} \log^{1/2}(N)$ corresponds to the bound $b_{\max}$ in (3.5). Although this bound depends on N, Theorem 3.3.2 can be applied, but to ensure the consistency of the classifier, we now have to prove that the estimation rate for $\widehat{b}_i$ decreases sufficiently fast.

Estimator of the diffusion coefficient. The construction of the estimator of σ^{*2} follows the same lines. However, since the diffusion coefficient is the same for all classes, we can use the whole dataset $\mathcal{D}_N$ to build its estimator with $A = \tilde{A}_N$, $\Xi = \tilde{K}_N$ and $(\tilde{A}_N)$, $(\tilde{K}_N)$ are increasing sequences of N. More precisely, we define

$$\tilde{\sigma}^2 \in \operatorname{argmin} h \in \mathcal{S}_{\tilde{K}_N,M} \frac{1}{nN} \sum_{j=1}^{N} \sum_{k=0}^{n-1} \left(U^j_{k\Delta_n} - h(X^j_{k\Delta_n})\right)^2, \quad \text{with} \quad U^j_{k\Delta_n} = \frac{(X^j_{(k+1)\Delta_n} - X^j_{k\Delta_n})^2}{\Delta_n}. \tag{3.9}$$

Finally, as for the drift estimator we consider the truncated version $\widehat{\sigma}^2$ as

$$\widehat{\sigma}^2(x) := \tilde{\sigma}^2(x)\mathbb{1}_{\{a_N \le \tilde{\sigma}^2(x) \le \tilde{A}_N \log^{1/2}(N)\}} + \tilde{A}_N \log^{1/2}(N)\mathbb{1}_{\{\tilde{\sigma}^2(x) > \tilde{A}_N \log^{1/2}(N)\}} + a_N \mathbb{1}_{\{\tilde{\sigma}^2(x) \le a_N\}}, \tag{3.10}$$

with $a_N = 1/\log(N)$. Although this constraint does not appear in Theorem 3.3.2, it remains natural in view of Assumption 3.2.1 (ii). We will impose that $\widehat{\sigma}^2$ is bounded by $\tilde{A}_N \log^{1/2}(N)$ to derive its consistency.

3.3.3 A general consistency result

In this section, we establish the consistency of the empirical classifier based on the estimators presented in the previous section. We first provide rates of convergence for the estimators of both the drift and the diffusion coefficients.

Theorem 3.3.4. *Let $i \in \mathcal{Y}$, and set $A_{N_i} = \log^2(N_i)$ conditional on the event $\{N_i > 1\}$, and $\tilde{A}_N = \log^2(N)$. Assume that Assumptions 3.2.1, 3.2.2 are satisfied. Considering the estimator $\widehat{b}_i$ of b_i^* (3.8) and the estimator $\widehat{\sigma}^2$ of σ^{*2} (3.10), set $K_{N_i} \propto (N_i \log(N_i))^{1/5}$ for $\widehat{b}_i$, and $\tilde{K}_N \propto (N \log(N))^{1/5}$ for $\widehat{\sigma}^2$. For N, N_i then n large enough, such that $\Delta_n = O(1/N)$, we have*

$$\mathbb{E}\left[\|\widehat{b}_i - b_i^*\|_{n,i}\right] \le C_1 \left(\frac{\log^4(N)}{N}\right)^{1/5}, \quad \text{and} \quad \mathbb{E}\left[\|\widehat{\sigma}^2 - \sigma^{*2}\|_n\right] \le C_2 \left(\frac{\log^4(N)}{N}\right)^{1/5},$$

where $C_1, C_2 > 0$ are constants which depend on L_0, $\mathfrak{p}_0$, and K.

Considering the estimation of the drift functions b_i^*, when $n, N_i \to \infty$ $a.s.$, we control the risk

$$\mathbb{E}\left[\|\widehat{b}_i - b_i^*\|_{n,N_i}^2\right] = \mathbb{E}\left[\frac{1}{nN}\sum_{j=j_1}^{j_{N_i}}\sum_{k=0}^{n-1}(\widehat{b}_i - b_i^*)^2(X_{k\Delta_n}^{(j)})\right],$$

using Equation (3.7), the properties of the constrained subspace $\mathcal{S}_{K_{N_i},M}$ together with that of functions of the spline basis (see proof of Theorem 3.3.4 for more details). Then, the control of the risk

$$\mathbb{E}\left[\|\widehat{b}_i - b_i^*\|_{n,i}^2\right] := \mathbb{E}\left[\sum_{k=0}^{n-1}(\widehat{b}_i - b_i^*)^2(X_{k\Delta_n})\right],$$

is obtained from the previous risk $\mathbb{E}\left[\|\widehat{b}_i - b_i^*\|_{n,N_i}^2\right]$ using, for all function h such that $\|h\|_\infty = \mathrm{O}(L_N)$ with (L_N) an increasing sequence of N , the following result

$$\mathbb{E}\left[\|h\|_n^2\right] - 2\mathbb{E}\left[\|h\|_{n,N}^2\right] = \mathrm{O}\left(\frac{K^* L_N \log(N)}{N}\right),$$

proved in [30], *Lemma A.2*.

Several comments can be made. First, we obtain a general rate of convergence for the estimation on $\mathbb{R}$ for both drift and diffusion coefficient functions under mild assumptions. This rate is, up to a logarithmic factor, of order $N^{-1/5}$. Hence, it extends the result of Theorem 3.3 in [30], where only consistency of drift estimators is obtained. In particular, a difficulty in establishing the convergence rate on $\mathbb{R}$ is to control the exit probabilities from the intervals $(-A_{N_i}, A_{N_i})$ and $(-\tilde{A}_N, \tilde{A}_N)$, which are provided here by careful estimates for the transition densities following [52].

This result together with Theorem 3.3.2 yields the consistency of the plug-in classifier

$$\widehat{g} := \widehat{g}_{\widehat{\mathbf{p}},\widehat{\mathbf{b}},\widehat{\sigma}^2}, \tag{3.11}$$

where the unknown parameters are replaced by their estimators in Equation (3.4). However, application of Theorem 3.3.2 requires the consistency of the estimator $\widehat{b}_i$ in terms of empirical norm $\|.\|_n$ and not in terms of norm $\|.\|_{n,i}$. To circumvent this issue, we can use a change of probability to get rid of the conditioning on $Y = i$. For this purpose, we take advantage of Lemma 3.7.3 and 3.7.4 to derive precise control of the transition density of the process X conditioned on $Y = i$, and then to establish the consistency of the plug-in classifier.

Theorem 3.3.5. *Grant Assumptions 3.2.1, 3.2.2. For N large enough, set $\Delta_n = O(1/N)$, $\tilde{A}_N = \log(N)$ and $\tilde{K}_N = (N\log(N))^{1/5}$. Moreover, for each $i \in \mathcal{Y}$, on the event $\{N_i > 1\}$, $A_{N_i} = \log(N_i)$ and $K_{N_i} \propto (N_i \log(N_i))^{1/5}$. Then, the classifier $\widehat{g}$ satisfies*

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \underset{N\to\infty}{\longrightarrow} 0.$$

The consistency of our classification procedure is obtained under very mild assumptions. The study of the rates of convergence requires more structural assumptions. In the following section, we obtain rates of convergence of the plug-in classifier under different kind of assumptions.

3.3.4 General rate of convergence for bounded drift function

In this section, we study the general rate of convergence of the proposed method described in Section 3.3.1 under the additional assumption that the drift functions of the considered mixture model are bounded, no additional assumptions being made on the diffusion coefficient.

Let us consider the following assumption.

Assumption 3.3.6. *There exists $C_{\mathbf{b}^*}$ such that*

$$\max_{i\in\mathcal{Y}} \|b_i^*\|_\infty \leq C_{\mathbf{b}^*}.$$

Let $i, j \in \mathcal{Y}^2$ with $i \neq j$. The following property allows to upper bound the expectation conditional on $\{Y = i\}$ by the expectation conditional on $\{Y = j\}$. This happens to be the cornerstone to derive rates of convergence for our procedure.

Proposition 3.3.7. *Under Assumptions 3.2.1, 3.2.2, and 3.3.6, we have for all $i, j \in \mathcal{Y}^2$ such that $i \neq j$, and N large enough*

$$\left\|\widehat{b}_i - b_i^*\right\|_{n,j}^2 \leq C \exp\left(\sqrt{c\log(N)}\right) \left\|\widehat{b}_i - b_i^*\right\|_{n,i}^2 + C\frac{A^2\log(N)}{N},$$

where $C, c > 0$ depend on $C_{\mathbf{b}^*}$, σ_1, and σ_0.

A crucial consequence of this result is that in particular the empirical norms $\|.\|_{n,i}$, $i \in \mathcal{Y}$, are now equivalent up to a factor of order $\exp\left(\sqrt{c\log(N)}\right)$. Notice that for all $r_1, r_2 > 0$,

$$\log^{r_1}(N) = o\big(\exp\left(\sqrt{c\log(N)}\right)\big), \quad \text{and} \quad \exp\left(\sqrt{c\log(N)}\right) = o\big(N^{r_2}\big). \tag{3.12}$$

In particular, the factor $\exp\left(\sqrt{c\log(N)}\right)$ is negligible with respect to any power of N. Therefore, combining Theorem 3.3.2, 3.3.4, and Proposition 3.3.7, we are able to give the rate of convergence for our procedure (when the drift coefficients are globally Lipschitz and bounded).

Theorem 3.3.8. *Grant Assumptions 3.2.1, 3.2.2, and 3.3.6. Set $\tilde{A}_N = \log(N)$ and $\tilde{K}_N \propto (N\log(N))^{1/5}$. Moreover, for each class $i \in \mathcal{Y}$, on the event $\{N_i > 1\}$, $A_{N_i} = \log(N_i)$ and $K_{N_i} \propto (N_i\log(N_i))^{1/5}$. The plug-in classifier $\widehat{g}$ given in Equation (3.11), provided that $\Delta_n = O\left(N^{-1}\right)$ and N large enough, satisfies*

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C \exp\left(\sqrt{c\log(N)}\right) N^{-1/5},$$

where $C > 0$ depends on $C_{\mathbf{b}^*}$, σ_1, and σ_0.

Leveraging the result of Theorem 3.3.4 and Proposition 3.3.7, we obtain a rate of convergence which is of order $N^{-1/5}$ up to the extra factor $\exp\left(\sqrt{c\log(N)}\right)$. Note that the optimal rate of convergence obtained when the estimation of drift function is done over on a compact set is of order $N^{-1/3}$ *w.r.t.* $\|.\|_n$ rather than $N^{-1/5}$ (see [30]). Here, this slower rate is mainly due to the fact that our procedure requires a control of the drift estimators over $\mathbb{R}$.

In the next section, we show that when σ^* is constant and assumed to be known, we derive faster rates of convergence. In particular, under Assumption 3.3.6, we show that our plug-in procedure achieves a rate of convergence of order $N^{-1/3}$. Lastly, note that Theorem 3.3.8 can be easily extended to higher order of regularity for the drift functions (*e.g.* Hölder with regularity $\beta > 1$). In this case, the obtained rate of convergence is of order $N^{-\beta/(2\beta+3)}$.

3.4 Classifier's rate of convergence with known diffusion coefficient

In this section, we consider that the diffusion coefficient is known and constant, and we derive faster rates of convergence of the classification procedure. For sake of simplicity, we choose $\sigma^* = 1$. In this case, our plug-in procedure only involves the estimation of the drift function $\widehat{\boldsymbol{b}}$. Hence, the plug-in classifier now writes as $\widehat{g} = \widehat{g}_{\mathbf{p},\widehat{\mathbf{b}},1}$.

In order to derive a general rate of convergence as a function of the drift regularity, we consider the following smoothness assumption [89], which is a subset of Lipschitz functions.

Assumption 3.4.1. *For all $i \in \mathcal{Y}$, b_i^* is Hölder with regularity parameter $\beta \geq 1$.*

3.4.1 Rates of convergence for drift estimators

Let $i \in \mathcal{Y}$. The study of the rates of convergence of the estimator $\widehat{b}_i$ relies on the properties of the matrix $\Psi_{K_{N_i}} \in \mathbb{R}^{(K_{N_i}+M)^2}$ defined by

$$\Psi_{K_{N_i}} := \left(\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}_{X|Y=i}\left[B_\ell(X_{k\Delta}^i)B_{\ell'}(X_{k\Delta}^i)\right]\right)_{\ell,\ell' \in [-M, K_{N_i}-1]}. \tag{3.13}$$

Note that for $t \in S_{K_{N_i},M},\ t = \sum_{i=-M}^{K_{N_i}-1} a_i B_{i,M,\mathbf{u}}$, we have the relation

$$\|t\|_{n,i}^2 = \mathbf{a}'\Psi_{K_{N_i}}\mathbf{a}, \quad \text{with } \mathbf{a} = \left(a_{-M}, \cdots, a_{K_{N_i}-1}\right)'.$$

Let us remind the reader that for a matrix P, the operator norm $\|P\|_{\text{op}}$ is defined as the square root of the largest eigenvalue of the matrix $P'P$. Besides, if P is symmetric, its norm is equal to its largest eigenvalue. The matrix $\Psi_{K_{N_i}}$ satisfies the following property.

Lemma 3.4.2. *Conditional on $(\mathbb{1}_{\{Y_1=i\}}, \ldots, \mathbb{1}_{\{Y_N=i\}})$, on the event $\{N_i > 1\}$, the matrix $\Psi_{K_{N_i}}$ given in Equation (3.13) satisfies*

(i) if $K_{N_i} \geq 1$, $\Psi_{K_{N_i}}$ is invertible,

(ii) under Assumption 3.2.1, for N large enough, if $K_{N_i} \leq \sqrt{N_i}$, there exists two constants $C, c > 0$ such that

$$c\frac{K_{N_i}}{A_{N_i}} \exp\left(\frac{A_{N_i}^2}{6}\right) \leq \|\Psi_{K_{N_i}}^{-1}\|_{\text{op}} \leq C\frac{K_{N_i}\log(N_i)}{A_{N_i}} \exp\left(\frac{2}{3}A_{N_i}^2\right).$$

A major consequence of Lemma 3.4.2 is to give the order of A_{N_i} *w.r.t.* N_i to obtain optimal rates of convergence for the estimation of the drift function b_i^*.

For fixed n and N_i in $\mathbb{N}^*$, let us denote

$$\Omega_{n,N_i,K_{N_i}} := \bigcap_{h \in \mathcal{S}_{K_{N_i},M}\setminus\{0\}} \left\{\left|\frac{\|h\|_{n,N_i}^2}{\|h\|_{n,i}^2} - 1\right| \leq \frac{1}{2}\right\}.$$

As we can see, the empirical norms $\|h\|_{n,N_i}$ and $\|h\|_{n,i}$ of any function $h \in \mathcal{S}_{K_{N_i},M} \setminus \{0\}$ are equivalent on the random set $\Omega_{n,N_i,K_{N_i}}$. More precisely, on the set $\Omega_{n,N_i,K_{N_i}}$, for all $h \in \mathcal{S}_{K_{N_i},M} \setminus \{0\}$, we have

$$\frac{1}{2}\|h\|_{n,i}^2 \leq \|h\|_{n,N_i}^2 \leq \frac{3}{2}\|h\|_{n,i}^2.$$

We use the random set $\Omega_{n,N_i,K_{N_i}}$ to derive a faster rate of the risk $\mathbb{E}\left[\|\widehat{b}_i - b^*_{A_{N_i},i}\|_{n,N_i}^2\right]$ of the estimator $\widehat{b}_i$ of the restriction $b^*_{A_{N_i},i}$ of the drift function b_i^* on the interval $[-A_{N_i}, A_{N_i}]$, with $A_{N_i} > 0$ on the event $\{N_i > 1\}$. We have for N large enough

$$\begin{aligned}\mathbb{E}\left[\|\widehat{b}_i - b^*_{A_{N_i},i}\|_{n,N_i}^2\right] &= \mathbb{E}\left[\|\widehat{b}_i - b^*_{A_{N_i},i}\|_{n,N_i}^2 \mathbb{1}_{\Omega_{n,N_i,K_{N_i}}}\right] + \mathbb{E}\left[\|\widehat{b}_i - b^*_{A_{N_i},i}\|_{n,N_i}^2 \mathbb{1}_{\Omega^c_{n,N_i,K_{N_i}}}\right] \\ &\leq \mathbb{E}\left[\|\widehat{b}_i - b^*_{A_{N_i},i}\|_{n,N_i}^2 \mathbb{1}_{\Omega_{n,N_i,K_{N_i}}}\right] + 4A_N^2\log(N)\mathbb{P}\left(\Omega^c_{n,N_i,K_{N_i}}\right),\end{aligned}$$

and the probability $\mathbb{P}\left(\Omega^c_{n,N_i,K_{N_i}}\right)$ satisfies

$$\mathbb{P}\left(\Omega^c_{n,N_i,K_{N_i}}\right) \leq 2(K_{N_i}+M)\exp\left(-C\frac{N_i}{A_{N_i}\left\|\Psi_{K_{N_i}}^{-1}\right\|_{\text{op}}}\right), \tag{3.14}$$

(see proof of Lemma 2.8.9, Equation (3.77) in Appendix). From Equation (3.14) and Lemma 3.4.2, we obtain

$$\mathbb{P}\left(\Omega^c_{n,N_i,K_{N_i}}\right) \leq 2(K_{N_i}+M)\exp\left(-C\frac{N_i}{K_{N_i}\log(N_i)}\exp\left(-\frac{2}{3}A_{N_i}^2\right)\right). \tag{3.15}$$

Notably, conditional on $(\mathbb{1}_{\{Y_1=i\}}, \ldots, \mathbb{1}_{\{Y_N=i\}})$ and on the event $\{N_i > 1\}$, if K_{N_i} is of order $N_i^{1/(2\beta+1)}$ (up to some extra logarithmic factors), and A_{N_i} is chosen such that the upper-bound of $\mathbb{P}\left(\Omega^c_{n,N_i,K_{N_i}}\right)$ is dominated by K_{N_i}/N_i as N tends to infinity, then the drift estimator converges as $N_i^{-2\beta/(2\beta+1)}$ *w.r.t.* $\|.\|_{n,i}^2$. Interestingly, this is the same rate of convergence obtained in [30] when the estimation

of the drift function is performed over a fixed compact interval. From this remark, if K_{N_i} is of order $\log^{-5/2}(N_i)N_i^{1/(2\beta+1)}$, and $A_{N_i} \leq \sqrt{\frac{3\beta}{2\beta+1}\log(N_i)}$, we deduce from Equation (3.15) that there exists a constant $C > 0$ such that

$$\mathbb{P}\left(\Omega^c_{n,N_i,K_{N_i}}\right) \leq 2(K_{N_i}+M)\exp\left(-C\log^{3/2}(N_i)\right),$$

and the desired result is obtained since $N_i \to \infty$ a.s. as $N \to \infty$. Furthermore, the lemma shows that the order of A_{N_i} is tight. Indeed, for another choice of A_{N_i} such that

$$\frac{A_{N_i}}{\sqrt{\log(N_i)}} \longrightarrow +\infty \ \text{ as } \ N \to +\infty,$$

then, from Equation (3.15), the upper-bound of $\mathbb{P}\left(\Omega^c_{n,N_i,K_{N_i}}\right)$ is of order K_{N_i} since

$$\exp\left(-C\frac{N_i}{K_{N_i}\log(N_i)}\exp\left(-\frac{2}{3}A^2_{N_i}\right)\right) \longrightarrow 1 \ \ a.s. \ \text{ as } N \to \infty,$$

and the convergence of $\mathbb{P}\left(\Omega^c_{n,N_i,K_{N_i}}\right)$ toward 0 is no longer guaranteed. Based on this observation, the next result establishes the rates of convergence for our proposed drift estimator on the event $\{N_i > 1\}$.

Theorem 3.4.3. *Let Assumptions 3.2.1, 3.2.2 and 3.4.1 be satisfied. Let $b^*_{A_{N_i},i} = b^*_i \mathbb{1}_{[-A_{N_i},A_{N_i}]}$ defined on the event $\{N_i > 1\}$. If $A_{N_i} \leq \sqrt{\frac{3\beta}{2\beta+1}\log(N_i)}$, $K_{N_i} \propto \left(\log^{-5/2}(N_i)N_i^{1/(2\beta+1)}\right)$, and $\Delta_n = O\left(N^{-1}\right)$. Then for all $i \in \mathcal{Y}$*

$$\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\mathbb{1}_{N_i>1}\right] \leq C\log^{6\beta}(N)N^{-2\beta/(2\beta+1)},$$

where C is a constant which depends on $\mathbf{b}^*$.

The above result shows that for a proper choice of A_{N_i} the drift estimators $\widehat{b}_i$ achieves, up to a logarithmic factor, the minimax rates of convergence *w.r.t.* $\|.\|_{n,i}$ (see Theorem 4.7 in [30]). Notably, Theorem 3.4.3 extends results obtained in [30] to the estimation of the drift function on a interval which depends on N.

In Section 3.4.2 and Section 3.4.3, we exploit this result to derive rates of convergence for the plug-in classifier $\widehat{g}$ defined as follows. On the event $\{\min_{i\in\mathcal{Y}} N_i > 1\}$, we consider the estimators $\widehat{\mathbf{b}}$ presented in Section 3.3.2, and define the plug-in classifier $\widehat{g} = \widehat{g}_{\widehat{\mathfrak{p}},\widehat{\mathbf{b}},1}$. On the complementary event $\{\min_{i\in\mathcal{Y}} N_i \leq 1\}$, we simply set $\widehat{g} = 1$.

3.4.2 Rates of convergence: bounded drift functions

In this section, we assume that additionally to $\sigma^* = 1$, Assumption 3.3.6 is fulfilled (the drift function is bounded). Hence, we can use Proposition 3.3.7, and apply Theorem 3.4.3 to derive rates of convergence for plug-in estimator $\hat{g}$.

Theorem 3.4.4. *Grant Assumptions 3.2.1, 3.2.2, 3.3.6, 3.4.1. Assume that for all $i \in \mathcal{Y}$, on the event $\{N_i > 1\}$, $A_{N_i} = \sqrt{\frac{6\beta}{2\beta+1}\log(N_i)}$ and $K_{N_i} \propto \left(\log^{-5/2}(N_i)N_i^{1/(2\beta+1)}\right)$, and $\Delta_n = O\left(N^{-1}\right)$. Then the plug-in classifier $\widehat{g} = \widehat{g}_{\widehat{\mathbf{b}},1}$ satisfies*

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\exp\left(\sqrt{c\log(N)}\right)N^{-\beta/(2\beta+1)},$$

where $C, c > 0$ are constants depending on $\mathbf{b}^*$, β, K and $\mathfrak{p}_0$.

The above theorem shows that the plug-in classifier $\widehat{g}$ achieves faster rates of convergence than in the case where σ^* is unknown (see Theorem 3.3.8). Notably, the obtained rate is of the same order,

up to a factor of order $\exp\left(\sqrt{c\log(N)}\right)$, than the rates of convergence provided in [44] in the framework of binary classification of functional data where the observation are assumed to come from a white noise model. In their setting, $\sigma^* = 1$ and the drift functions depend only on the observation time interval, which is also assumed to be $[0,1]$. Therefore, our specific setup is more challenging since the drift functions are space-dependent which involves to deal with estimation of function on a non-compact interval. Finally, it is worth noting that, up to the $\exp\left(\sqrt{c\log(N)}\right)$ factor, the rate of convergence provided in Theorem 3.4.4 is the same as the minimax rates in the classical classification framework where the feature vector X belongs to $\mathbb{R}^d$ and that X admits a lower bounded density [95, 4].

3.4.3 Rates of convergence: when the drift functions are re-entrant

In this section, we study performance of the plug-in classifier when the drift functions are not necessarily bounded. In this context, rates of convergence are obtained under the following assumption.

Assumption 3.4.5. *(re-entrant drift function) For each label $i \in \mathcal{Y}$, there exists $c_0 > 4$ and $K_0 \in \mathbb{R}$ such that*

$$\forall x \in \mathbb{R},\ \ b_i^*(x)x \leq -c_0 x^2 + K_0.$$

An important consequence of this assumption is that there exists $C > 0$ (see Proposition 1.1 in [52]) such that

$$\mathbb{E}\left[\exp(4|X_t|^2)\right] \leq C, \tag{3.16}$$

which yields a better bound on the tail probability $\mathbb{P}\left(|X_t| \geq A\right)$ for $A > 0$. It worth noting that under Assumption 3.4.5, the drift functions are not bounded. Hence, we can not take advantage of Proposition 3.3.7 to derive rates of convergence. Nonetheless, we obtain the following result.

Theorem 3.4.6. *Grant Assumptions 3.2.1, 3.2.2, 3.4.1, 3.4.5. Assume that for all $i \in \mathcal{Y}$, on the event $\{N_i > 1\}$, $A_{N_i} = \sqrt{\frac{3\beta}{2\beta+1}\log(N_i)}$ and $K_{N_i} \propto \left(\log^{-5/2}(N_i)N_i^{1/(2\beta+1)}\right)$, and $\Delta_n = O\left(N^{-1}\right)$. Then, the plug-in classifier $\widehat{g}$ satisfies*

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\log^{3\beta+1}(N)N^{-3\beta/4(2\beta+1)}.$$

The above theorem shows that the rate of convergence of the plug-in classifier is, up to a logarithmic factor, of order $N^{-3\beta/4(2\beta+1)}$. Therefore, this rate of convergence is slightly slower than the one provided in Theorem 3.4.4. It is mainly due to the fact that under Assumption 3.4.5, Proposition 3.3.7 does not apply and then, in view of considered assumptions in Theorem 3.4.6, we only manage to obtain the following bound,

$$\forall i,j \in \mathcal{Y}:\ i \neq j,\ \mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,j}^2\right] \leq CN^{\beta/4(2\beta+1)}\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,i}^2\right],$$

which is clearly worst than the one obtained in Proposition 3.3.7. Interestingly, for $\beta = 1$, we can note that the rates obtained in Theorem 3.4.6 are of the same order than the rates of convergence established in [45] in the classification setup where the input vector lies in $\mathbb{R}$ under the assumption that X does not fulfil the strong density assumption (*e.g.* the density of X is not lower bounded).

3.5 Simulation study

This section is devoted to numerical experiments that support our theoretical findings. A first part is dedicated to the study of the performance of the plug-in classifier in a setting which meets the assumptions of Section 3.3.4. The considered model is presented in Section 3.5.1. The implementation of the proposed procedure is discussed in Section 3.5.2 while the performances of the plug-in classifier are given in Section 3.5.3. Finally, several features of the problem are investigated in Section 3.5.4. In particular, we consider the classical Ornstein-Uhlenbeck model, for which assumptions of Section 3.3.4 are not fulfilled.

3.5.1 Models and simulation setting

We fix $K = 3$ classes in the following. Note that, we do not consider larger value of K since the evaluation of the impact of K on the procedure is beyond the scope of this paper. To illustrate the accuracy of the presented plug-in classifier, we investigate the model described in Table 3.1. This

$b_1^*(x)$	$1/4 + (3/4)\cos^2 x$
$b_2^*(x)$	$\theta[1/4 + (3/4)\cos^2 x]$
$b_3^*(x)$	$-\theta[1/4 + (3/4)\cos^2 x]$
$\sigma^*(x)$	$0.1 + 0.9/\sqrt{1+x^2}$

Table 3.1: *Drift and diffusion coefficients, depending on* $\theta \in \Theta = \{1/2, 3/4, (4+\alpha)/4, \alpha \in [\![1, 12]\!]$.

toy model, described in Table 3.1, fulfills the assumptions of Section 3.3.4. Interestingly, this model allows evaluating the influence of the distance between the drift functions of each of the three classes, on the classification problem, through the parameter θ. Indeed,

$$\min_{i,j=1,2,3} \|b_i^* - b_j^*\|_\infty = \theta, \quad \text{where } \theta \in \Theta = \{1/2, 3/4, (4+\alpha)/4, \alpha \in [\![1, 12]\!]\}.$$

We investigate the consistency of the empirical classifier using learning samples of size $N \in \{100, 1000\}$ with $n \in \{100, 500\}$ (and thus with $\Delta_n = 1/n$). We use the `R`-package `sde` (see [59]) to simulate the solution of the stochastic differential equation corresponding to the chosen model.

Figure 3.1 displays simulated trajectories from the proposed model. On the left panel (right panel respectively) the observed learning sample comes from the model with parameter $\theta = 1/2$ ($\theta = 4$ respectively) and each class is represented by one color. We can see from Figure 3.1 that the distance between the drift functions strongly impacts the dispersion of the trajectories and leads to a more difficult classification task.

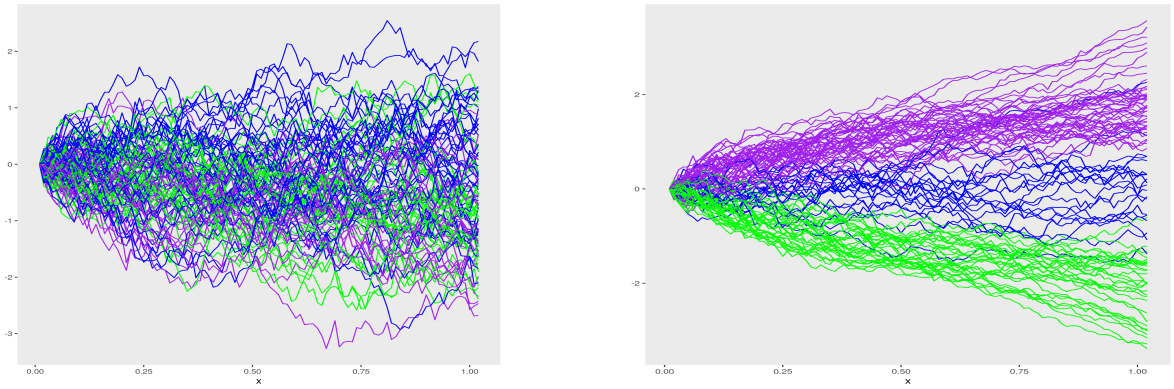

Figure 3.1: *Dispersion of diffusion paths from model given in Table 3.1. Left:* $\theta = 1/2$*, right:* $\theta = 4$ *(blue lines* $K = 1$*, purple lines* $K = 2$*, green lines* $K = 3$*); with* $N = 100$ *and* $n = 100$*.*

Performance of the Bayes classifier. We evaluate the performance of the Bayes classifier g^* with respect to four values of parameter θ ($\theta \in \{1/2, 3/2, 5/2, 4\}$). To this end, we compute its average error rate over 100 repetitions of the following steps

(i) simulate $\mathcal{D}_N$ of size $N = 4000$ with $n = 500$;

(ii) based on $\mathcal{D}_N$ compute the misclassification error rate of the discrete counterpart of g^*.

Table 3.2 provides the mean and standard deviation of the misclassification risk. The obtained results highlight the significant impact of the minimum distance θ, between the drift functions of each class, on the performance of g^*. Indeed, as expected, the Bayes classifier is more accurate on our model

when parameter θ is large, especially in the case of separable data ($\theta = 4$). On the contrary, the worst case corresponds to $\theta = 0.5$. In this model, the data are highly ambiguous.

	$\theta = 1/2$	$\theta = 3/2$	$\theta = 5/2$	$\theta = 4$
$\widehat{\mathcal{R}}(g^*)$	0.49 (0.01)	0.36 (0.01)	0.22 (0.01)	0.11 (0.01)

Table 3.2: *Classification risks of the Bayes classifier g^* w.r.t parameter θ from learning samples of size $N = 4000$ with $n = 500$.*

3.5.2 Implementation of the plug-in classifier

Hereafter, we briefly describe the implementation of the proposed plug-in classifier. We first estimate the drift functions b_i^*, $i = 1, 2, 3$. For each $i \in \{1, 2, 3\}$, the estimator $\widehat{b}_i$ is built on the interval $[-A_{N_i}, A_{N_i}]$. Since the drifts (and the diffusion) coefficients are bounded, we can use the construction considered in Section 3.3. Therefore, we fix $A_{N_i} = \log(N_i)$, $M = 3$, and divide the learning sample $\mathcal{D}_N$ into sub-samples $\mathcal{D}_N^i$ of size N_i that contains all diffusion paths belonging to the class i. From the sub-sample $\mathcal{D}_N^i$, we build estimators $\widehat{b}_i$, $i = 1, 2, 3$.

For the construction of the estimator $\widehat{b}_i$, we have to choose the dimension parameter K_{N_i}. We follow [30], and consider an adaptive choice denoted by $\widehat{K}_{N_i}$.

Let us remind the reader that in [30], the adaptive dimension $\widehat{K}_{N_i}$ is selected such that $\widehat{K}_{N_i}$ is the minimizer of the following penalized contrast

$$\widehat{K}_{N_i} := \underset{K \in \mathcal{K}}{\operatorname{argmin}} \left\{ \frac{1}{Nn} \sum_{j=1}^{N} \sum_{k=0}^{n-1} (\widehat{b}_{i,K} - Z^j_{k\Delta_n})^2 + \operatorname{pen}_b(K) \right\}, \tag{3.17}$$

where $\mathcal{K} \in \{2^q,\ q \in [\![0, 5]\!]\}$, and $\widehat{b}_{i,K}$ is the drift estimator built on the approximation subspace $\mathcal{S}_{K,M}$. Besides, $\operatorname{pen}_b(K) = \kappa(K + M)\log^3(N)/N$ is the penalty function with $\kappa > 0$. We fix the parameter $\kappa = 0.1$ as recommended in [30]. For the estimation of σ^2, we consider the whole sample $\mathcal{D}_N$ and apply the methodology described in Section 3.3 with $M = 3$. We follow the same lines to build an adaptive estimator of σ^{2*}, and choose $\widehat{K}_N$ as the minimizer over $\mathcal{K}$ of the following penalized contrast

$$\widehat{K}_N := \operatorname{argmin} K \in \mathcal{K} \left\{ \frac{1}{Nn} \sum_{j=1}^{N} \sum_{k=0}^{n-1} (\widehat{\sigma}^2_K - U^j_{k\Delta_n})^2 + \operatorname{pen}_\sigma(K) \right\}, \tag{3.18}$$

where $\hat{\sigma}^2_K$ is the estimator built on $\mathcal{S}_{K,M}$, and $\operatorname{pen}_\sigma(K) := \kappa(K + M)\log^3(N)/Nn$ is the penalty function, with $\kappa > 0$. The value of the tuning parameter κ is calibrated through an intensive simulation study and chosen equal to $\kappa = 5$.

The function `SDEclassif` of the `R`-package SDEclassif, available on github, implements the resulting plug-in classifier.

3.5.3 Simulation results

The performance of the plug-in classifier $\widehat{g}$ is evaluated by repeating 100 times the following steps

1. Simulate learning samples $\mathcal{D}_N$ and $\mathcal{D}_{N'}$ with $N \in \{100, 1000\}$, $N' = 1000$, and $n \in \{100, 500\}$;
2. for each $i \in \{1, 2, 3\}$, from the sub-sample $D_N^i = \{\bar{X}^j, j \in \mathcal{I}_i\}$, select $\widehat{K}_N$ minimizing (3.17) and compute the estimator $\widehat{b}_{i,\widehat{K}_N}$ of b_i^* given in Equation (3.8);
3. from $\mathcal{D}_N$ select $\widehat{K}_N$ using Equation (3.18) and compute the estimator $\widehat{\sigma}^2_{\widehat{K}_N}$ of σ^{*2} given in (3.10);
4. based on $\mathcal{D}_N$ compute $\widehat{\mathfrak{p}} = \left(\frac{1}{N}\sum_{j=1}^N \mathbb{1}_{Y_j=1}, \frac{1}{N}\sum_{j=1}^N \mathbb{1}_{Y_j=2}, \frac{1}{N}\sum_{j=1}^N \mathbb{1}_{Y_j=3}\right)$;

5. based on $\mathcal{D}_{N'}$, compute the error rate of the plug-in classifier $\widehat{g}$ where $\widehat{\mathbf{b}} = \left(\widehat{b}_{1,\widehat{K}_N}, \widehat{b}_{2,\widehat{K}_N}, \widehat{b}_{3,\widehat{K}_N}\right)$ and $\widehat{\sigma}^2 = \widehat{\sigma}^2_{\widehat{K}_N}$, and $\widehat{\mathfrak{p}}$.

From these repetitions, we compute the empirical mean and standard deviation of the error rate of $\widehat{g}$. The results are given in Table 3.3 and Figure 3.2. As expected, from Table 3.3 and Table 3.2, we can see that the error rate of the plug-in classifier $\widehat{g}$ is closed to the error rate of the Bayes classifier. In particular, for $N = 1000$, it performs as well as the Bayes classifier. Note that the length of the paths n does not significantly impact the performance of $\widehat{g}$. Moreover, from Figure 3.2, we can make similar comments as for the Bayes classifier (see Table 3.2), in particular, the accuracy of $\widehat{g}$ decreases as parameter θ increases.

$\widehat{\mathcal{R}}(\widehat{g})$	$n = 100$		$n = 500$	
	$N = 100$	$N = 1000$	$N = 100$	$N = 1000$
$\theta = 1/2$	0.53 (0.05)	0.50 (0.05)	0.53 (0.05)	0.49 (0.05)
$\theta = 3/2$	0.39 (0.06)	0.37 (0.05)	0.39 (0.05)	0.36 (0.05)
$\theta = 5/2$	0.24 (0.05)	0.22 (0.04)	0.25 (0.04)	0.22 (0.04)
$\theta = 4$	0.12 (0.03)	0.10 (0.03)	0.11 (0.03)	0.10 (0.03)

Table 3.3: *Risks of the plug-in classifier $\widehat{g}$ w.r.t. values of parameter θ*

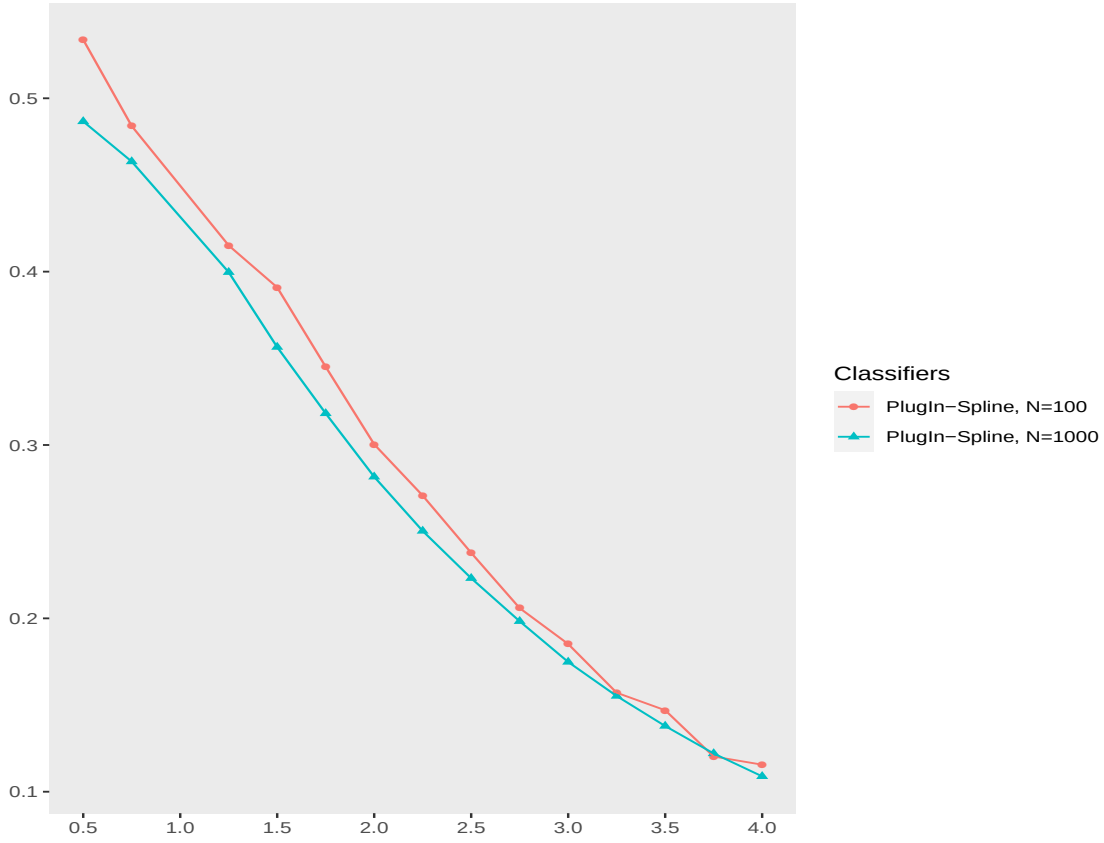

Figure 3.2: *Risks of the plug-in classifier w.r.t values of the minimum gap θ between the drift functions*

3.5.4 Ornstein-Uhlenbeck model

In this section, we focus on the influence of the diffusion coefficient σ^* on the performance of our plug-in procedure. To this end, we consider the Ornstein-Uhlenbeck diffusion model given in Table 3.4 where the diffusion coefficient σ^* is constant. Let us notice also that in this model the drift functions are unbounded. We investigate the performance of the plug-in classifier $\widehat{g}$ *w.r.t.* the level

$b_1^*(x)$	$1-x$
$b_2^*(x)$	$-1-x$
$b_3^*(x)$	$-x$
$\sigma^*(x)$	σ

Table 3.4: *Ornstein-Uhlenbeck mixture model with* $K=3$

of noise σ^*. This study is motivated by the fact that, inherently, the diffusion coefficient impacts the dispersion of the trajectories. Therefore, it can lead to separable data when σ^* is close to zero, and ambiguous data for large values of σ^*. Thus, we evaluate the performance of $\widehat{g}$ for $\sigma^*=1/2$ which is close enough to zero, and for larger value $\sigma^* \in \{1, 3/2\}$. We first consider the case where σ^* is unknown. The results are given in Table 3.5 and confirm our intuition. The error rate of the plug-in classifiers decreases as σ^* decreases.

In a second step, we investigate the influence of estimating the coefficient σ^* in the procedure. To evaluate this point, we assess the error rate of the plug-in classifier when $\sigma^*=1$ is known. In this case, we only estimate the drift functions and the weights of mixture $\mathfrak{p}p^*$ to build our predictor. The results are given in Table 3.6. First, we can notice that by comparison with results provided in Table 3.5, there is almost no impact on the performance of the plug-in classifier when we assume the diffusion coefficient σ^* in the Ornstein-Uhlenbeck model to be known or not.

Finally, we also study the influence of parameter A_N on the estimation procedure. Indeed, our theoretical results indicates that A_N should be of order $\sqrt{\log(N)}$ when σ^* is constant and known, while $A_N=\log(N)$ is recommended when σ^* is unknown. To this end, we evaluate the error rate of our procedure for these choices. The results are also provided in Table 3.6 and show that the performance are almost the same in the two cases.

	$\widehat{\mathcal{R}}(\widehat{g})$	$\widehat{\mathcal{R}}(g^*)$
$\sigma^*=1/2$	0.23 (0.04)	0.21 (0.01)
$\sigma^*=1$	0.44 (0.05)	0.41 (0.01)
$\sigma^*=3/2$	0.52 (0.05)	0.49 (0.01)

Table 3.5: *Evolution of the performance of the plug-in classifier* $\widehat{g}$ *and of* g^* *w.r.t values of the constant diffusion coefficient* σ^* *for* $N=100$ *and* $n=100$.

	$N=100$	$N=1000$
$A_N=\sqrt{\log(N)}$	0.44 (0.05)	0.41 (0.05)
$A_N=\log(N)$	0.43 (0.05)	0.43 (0.05)

Table 3.6: *Risk classification of* $\widehat{g}$ *when the diffusion* $\sigma^*=1$ *is known, and* $n=100$.

3.6 Conclusion and discussion

In this paper, we propose a plug-in classifier for the multiclass classification of trajectories generated by a mixture of diffusion processes whose drift functions $b_i^*,\ i \in \mathcal{Y}$ and diffusion coefficient σ^* are assumed to be unknown. In the considered model, each class i is characterized by a drift function, b_i^* whereas the diffusion coefficient σ^* is common for all classes. This work extends to the nonparametric case, the multiclass classification procedure provided in [29] where $\sigma^* = 1$ and the drift functions depend on an unknown parameter $\theta \in \mathbb{R}^d$. Our proposed procedure relies on consistent projection estimators $\widehat{b}_i, i \in \mathcal{Y}$ and $\widehat{\sigma}^2$ of the drift and diffusion coefficients on a constrained approximation subspace spanned by the spline basis. We establish the consistency, *w.r.t.* the excess risk, of our procedure and then studied its rate of convergence under different kind of assumptions. In particular, we show that the proposed plug-in classifier reaches a rate of convergence of order $N^{-1/5}$ (up to a factor of order $\exp(\sqrt{c\log(N)})$) when $\mathbf{b}^*,\ \sigma^*$, and $\mathfrak{p}^*$ are unknown. Besides, a numerical study illustrates the performance of our classification procedure.

In the case where $\sigma^* = 1$, we manage to derive faster rates of convergence. In particular, when the drift functions are bounded and Hölder with regularity $\beta \geq 1$, we obtained a rate of order $N^{-\beta/(2\beta+1)}$ (up to a factor of order $\exp(\sqrt{c\log(N)})$). Interestingly, this result can be viewed as an extension of the one obtained in [44] to the multiclass mixture model, where the drift functions are time-dependent. Furthermore, up to $\exp(\sqrt{c\log(N)})$ factor, our rate of convergence matches the optimal rates of convergence obtained in the univariate setting (*e.g.* $X \in \mathbb{R}$), in [4]. Finally, for the case of unbounded drift functions, we assume that the drift functions are the re-entrant. Taking advantage of this property, we establish that our plug-in classifier achieves a rate of convergence of order $N^{-3\beta/4(2\beta+1)}$. For $\beta = 1$, this rate of convergence is of the same order as the one obtained in [45] for plug-in classifier in the univariate classification setting, when the feature X does not satisfy the strong density assumption.

A question that can be tackled for future research is the study of the optimality in the minimax sense of our plug-in procedure. In particular, the adaptivity of estimators of the drift and diffusion coefficients should be investigated. Furthermore, it might be interesting to consider the margin type assumption as in [44] to derive faster rates of convergence. Also, following [29], it is natural to derive theoretical properties for empirical risk minimization procedure based on convex losses. Finally, the extension to the high-dimensional setting would require further work. In particular, the control of the transition densities is different in this setting.

3.7 Proofs

The section is devoted to the proofs of our main results. In order to simplify the notation, we write $\Delta_n = \Delta$. Besides, $C > 0$ is a constant which may change from one line to another. When the dependency on a parameter θ needs to be highlighted, we write C_θ.

3.7.1 Technical results on the process X

Lemma 3.7.1. *Under Assumption 3.2.1 and for all integer $q \geq 1$, there exists $C^* > 0$ depending on q such that for all $0 \leq s < t \leq 1$,*

$$\mathbb{E}\,|X_t - X_s|^{2q} \leq C^*(t-s)^q.$$

The proof of Lemma 3.7.1 is provided in Appendix.

For each $t \in [0,1]$ and $x \in \mathbb{R}$, we denote by $p_X(t,x)$ the transition density of the underlying process X_t given the starting point $X_0 = 0$. We also denote by $p_{i,X}(t,\cdot)$ the transition density of the process driven by the drift function b_i^*. Note that Assumption 3.2.1 ensures the existence of the transition densities. The rest of this section is dedicated to some results on the transition densities $p_{i,X}$ for $i = 1,\dots,K$. Nonetheless, since the transition p_X of the process X writes as

$$p_X = \sum_{i=1}^{K} \mathfrak{p}_i^* p_{i,X},$$

all these results apply also for p_X. The following proposition is provided in [52] (Proposition 1.2).

Proposition 3.7.2. *Under Assumptions 3.2.1 and 3.2.2, there exist constants $c > 1$, $C > 1$ such that for all $t \in (0,1]$, $x \in \mathbb{R}$, and $i = 1, \dots, K$*

$$\frac{1}{C\sqrt{t}} \exp\left(-c\frac{x^2}{t}\right) \leq p_{i,X}(t,x) \leq \frac{C}{\sqrt{t}} \exp\left(-\frac{x^2}{ct}\right).$$

From this result, we can deduce an evaluation of the probability of the process to exit a compact set. This is the purpose of the next result.

Lemma 3.7.3. *Under Assumption 3.2.1 and 3.2.2, there exist $C_1, C_2 > 0$ such that for all $A > 0$*

$$\sup_{t\in[0,1]} \mathbb{P}\left(|X_t| \geq A\right) \leq \frac{C_1}{A} \exp(-C_2 A^2).$$

Proof. Let $A > 0$, we have for $t \in (0,1]$,

$$\mathbb{P}\left(|X_t| \geq A\right) = \int_A^{+\infty} p_X(t,x)\mathrm{d}x + \int_A^{+\infty} p_X(t,-x)\mathrm{d}x.$$

From Proposition 3.7.2, we then deduce that

$$\mathbb{P}\left(|X_t| \geq A\right) \leq C\frac{\sqrt{t}}{A} \int_A^{+\infty} c\frac{2x}{t} \exp\left(-c\frac{x^2}{t}\right) \mathrm{d}x \leq \frac{C\sqrt{t}}{A} \exp\left(-\frac{cA^2}{t}\right).$$

From the above inequality, and using that $t \in (0,1]$, we deduce the result. □

Lemma 3.7.4. *Under Assumption 3.2.1, there exist C_0, C_1, and C_2, such that for $i = 1, \dots, K$, for $x \in [-A, A]$, we have*

$$C_1 \exp\left(-C_2 A^2\right) \leq \frac{1}{n} \sum_{k=1}^{n-1} p_{i,X}(k\Delta, x) \leq C_0.$$

Proof of Lemma 3.7.4. For $i \in \{1, \dots, K\}$, for all $x \in \mathbb{R}$, we have from Proposition 3.7.2,

$$\frac{1}{n}\sum_{k=1}^{n} p_{i,X}(k\Delta, x) \leq \frac{C}{n}\sum_{k=1}^{n} \frac{1}{\sqrt{k\Delta}} = \frac{C}{\sqrt{n}}\sum_{k=1}^{n}\frac{1}{\sqrt{k}} \leq \frac{2C}{\sqrt{n}}\sum_{k=1}^{n}\frac{1}{\sqrt{k+1}}. \tag{3.19}$$

Since the function $x \mapsto \frac{1}{\sqrt{x}}$ is decreasing over $[1, +\infty[$, we deduce from Equation (3.19) that

$$\frac{1}{n}\sum_{k=1}^{n} p_{i,X}(k\Delta, x) \leq \frac{4C\sqrt{n+1}}{\sqrt{n}} \leq C_0,$$

which gives the upper bound. For the lower bound, we observe from Proposition 3.7.2 that for $k \in [\![1, n-1]\!]$, and $x \in \mathbb{R}$,

$$C\exp\left(-\frac{cx^2}{k\Delta}\right) \leq \frac{C}{\sqrt{k\Delta}} \exp\left(-\frac{cx^2}{k\Delta}\right) \leq p_i(k\Delta, x). \tag{3.20}$$

Since $g : (s,x) \mapsto \exp\left(-\frac{cx^2}{s}\right)$ is strictly increasing in s over $(0,1]$, we obtain for $k \in [\![1, n-1]\!]$,

$$\int_{\frac{1}{n}}^{\frac{n-1}{n}} g(s,x)ds \leq \sum_{k=2}^{n-1} \int_{(k-1)\Delta}^{k\Delta} \left(g(k\Delta, x) + g(s,x) - g(k\Delta, x)\right) ds \leq \frac{1}{n}\sum_{k=1}^{n-1} \exp\left(-\frac{cx^2}{k\Delta}\right).$$

Hence, we deduce that for $n \geq 3$, and $x \in [-A, A]$

$$\frac{1}{6}\exp(-2cA^2) \leq \int_{\frac{1}{2}}^{\frac{n-1}{n}} g(s,x)\mathrm{d}s \leq \frac{1}{n}\sum_{k=1}^{n-1} \exp\left(-\frac{cx^2}{k\Delta}\right).$$

For the first lower bound, we use that $g(s,x) \geq e^{-2cA^2}$ for $x \in [-A, A]$ and $s \geq 1/2$, and that the length of $[1/2, (n-1)/n]$ is larger than $1/6$ for $n \geq 3$. This explains our choice of integration interval in the middle term of the above inequalities. Finally, gathering this bound with Equation (3.20), leads to

$$\frac{1}{6}\exp(-2cA^2) \leq \frac{1}{n}\sum_{k=1}^{n-1}\exp\left(-\frac{cx^2}{k\Delta}\right) \leq \frac{1}{n}\sum_{k=1}^{n-1} p_{i,X}(k\Delta, x).$$

□

Lemma 3.7.5. *Suppose that σ^* is a constant. Under Assumption 3.2.1, and for all $q > 1$, there exists $K_q > 1$ such that for all $(t,x) \in (0,1] \times [-A, A]$,*

$$\frac{1}{K_q\sqrt{t}}\exp\left(-\frac{2q-1}{2q\sigma^{*2}t}x^2\right) \leq p_X(t,x) \leq \frac{K_q}{\sqrt{t}}\exp\left(-\frac{x^2}{2q\sigma^{*2}t}\right).$$

Proof of Lemma 3.7.5. The transition density p_X^0 of the process $(0 + \sigma^* W_t)_{t\in]0,1]}$ (with a constant diffusion coefficient σ^*) is given by

$$p_X^0(t,x) := \frac{1}{\sqrt{2\pi\sigma^{*2}t}}\exp\left(-\frac{1}{2\sigma^{*2}t}|0-x|^2\right). \tag{3.21}$$

We are going to demonstrate the inequality for $p_{i,X}$, which is the transition density of X in class number i. Indeed, then it will be true for all $i \in \mathcal{Y}$ and thus for $p_X = p_{Y,X}$. We follow here the arguments given in the *proof of (1.6)* in [52]. Let us denote,

$$Z_{i,t} = \exp\left(\int_0^t \frac{b_i^*(X_s)}{\sigma^*}dW_s - \int_0^t \frac{b_i^{*2}(X_s)}{\sigma^{*2}}ds\right).$$

We have $\forall (t,x) \in]0,1] \times \mathbb{R}$,

$$p_{i,X}(t,x) = p_X^0(t,x)\mathbb{E}^0\left[Z_{i,t}|X_t = x\right],$$

and

$$\frac{1}{p_{i,X}(t,x)} \leq \frac{1}{p_X^0(t,x)}\mathbb{E}^0\left[Z_{i,t}^{-1}|X_t = x\right]. \tag{3.22}$$

Then,

$$\mathbb{E}^0\left[Z_{i,t}|X_t = x\right] = 1 + \frac{1}{p_X^0(t,x)}\int_0^t \mathbb{E}^0\left[Z_{i,t}b_i^*(X_s)\frac{X_s - x}{\sigma^{*2}(t-s)}p_X^0(t-s,x)\right]ds$$

and

$$\mathbb{E}^0\left[Z_{i,t}^{-1}|X_t = x\right] = 1 + \frac{1}{p_X^0(t,x)}\int_0^t \mathbb{E}^0\left[Z_{i,s}^{-1}b_i^*(X_s)\frac{X_s - x}{\sigma^{*2}(t-s)}p_X^0(t-s,x)\right]ds.$$

For all $(t,x) \in]0,1] \times \mathbb{R}$, one has :

$$\begin{aligned}
\mathbb{E}^0\left[Z_{i,t}|X_t = x\right] &= 1 + \frac{1}{p_X^0(t,x)}\int_0^t \mathbb{E}^0\left[Z_{i,s}b_i^*(X_s)\frac{X_s - x}{\sigma^{*2}(t-s)}p_X^0(t-s,x)\right]ds \\
&\leq 1 + \frac{C}{p_X^0(t,x)}\int_0^t \mathbb{E}^0\left[Z_{i,s}\left|b_i^*(X_s)\right|\frac{|X_s - x|}{(t-s)^{3/2}}\exp\left(-\frac{(X_s-x)^2}{2\sigma^{*2}(t-s)}\right)\right]ds \\
&\leq 1 + \frac{C}{p_X^0(t,x)}\int_0^t \mathbb{E}^0\left[Z_{i,s}\left|b_i^*(X_s)\right|\frac{1}{\varepsilon(t-s)}\exp\left(-\frac{(1-\varepsilon)(X_s-x)^2}{2\sigma^{*2}(t-s)}\right)\right]ds,
\end{aligned}$$

using that $y\varepsilon\exp(-\varepsilon y^2/2) \leq 1$ for $0 < \varepsilon < 1$. Let $q, q' > 1$ be two real numbers such that $\frac{1}{q} + \frac{1}{q'} = 1$. Using Hölder's inequality, and the Lipschitz property of b^*, one has:

$$\mathbb{E}^0\left[Z_{i,t}|X_t = x\right] \leq 1 + \frac{C\varepsilon^{-1}}{p_X^0(t,x)}\int_0^t \left(\mathbb{E}^0\left[\frac{Z_{i,s}^q q\left(1+|X_s|\right)^q}{(t-s)^q}\right]\right)^{\frac{1}{q}}\left(\mathbb{E}^0\left[\exp\left(-\frac{(X_s-x)^2}{2\sigma^{*2}(t-s)}\right)\right]\right)^{\frac{1}{q'}}ds, \tag{3.23}$$

with $\varepsilon = 1 - 1/q'$. According to *Lemma A.1 in [52]*, one has:

$$\forall q > 1, \;\; \mathbb{E}^0\left[Z_{i,s}^{q}\left(1+|X_s|\right)^q\right] + \mathbb{E}^0\left[Z_{i,s}^{-q}\left(1+|X_s|\right)^q\right] \leq C_1,$$

where $C_1 > 0$ is a constant. Thus, it remains to upper bound $\mathbb{E}^0\left[\exp\left(-\frac{(X_s-x)^2}{2\sigma^{*2}(t-s)}\right)\right]$ and then deduce an upper bound of $\mathbb{E}^0\left[Z_{i,t}|X_t = x\right]$. For all $s < t$, we have:

$$\sqrt{2\pi\sigma^{*2}s}\mathbb{E}^0\left[\exp\left(-\frac{(X_s-x)^2}{2\sigma^{*2}(t-s)}\right)\right] = \int_{\mathbb{R}} \exp\left(-\frac{1}{2\sigma^{*2}(t-s)}\left(z-x\right)^2\right)\exp\left(-\frac{1}{2\sigma^{*2}s}z^2\right)dz.$$

It follows that,

$$\mathbb{E}^0\left[\exp\left(-\frac{(X_s-x)^2}{2\sigma^{*2}(t-s)}\right)\right] = \sqrt{\frac{t-s}{t}}\exp\left(-\frac{x^2}{2\sigma^{*2}t}\right).$$

Thus, from Equation (3.23), we obtain:

$$\begin{aligned}
\mathbb{E}^0\left[Z_{i,t}|X_t = x\right] &\leq 1 + \frac{C\varepsilon^{-1}}{p_X^0(t,x)}\int_0^t \frac{(t-s)^{\frac{1}{2q'}-1}}{t^{\frac{1}{2q'}}}\exp\left(-\frac{x^2}{2q'\sigma^{*2}t}\right)ds \\
&\leq 1 + \frac{C\varepsilon^{-1}t^{-1/2q'}}{p_X^0(t,x)}\exp\left(-\frac{x^2}{2q'\sigma^{*2}t}\right)\left[-2q'(t-s)^{1/2q'}\right]_0^t \\
&\leq 1 + \frac{C\varepsilon^{-1}}{p_X^0(t,x)\sqrt{t}}\exp\left(-\frac{x^2}{2q'\sigma^{*2}t}\right),
\end{aligned}$$

From the definition of function p_X^0 given in Equation (3.21) together with relation (3.22), we obtain that

$$p_{i,X}(t,x) \leq p_X^0(t,x)\left(1 + \frac{C\varepsilon^{-1}}{p_X^0(t,x)\sqrt{t}}\exp\left(-\frac{x^2}{2q'\sigma^{*2}t}\right)\right).$$

Thus, there exists a constant $K_q > 1$ (as $\varepsilon = 1 - 1/q'$ and $1/q + 1/q' = 1$) such that,

$$\forall (t,x) \in]0,1]\times\mathbb{R}, \;\; p_{i,X}(t,x) \leq \frac{K_q}{\sqrt{t}}\exp\left(-\frac{x^2}{2q'\sigma^{*2}t}\right), \;\; \forall q' > 1. \tag{3.24}$$

Following the same lines, one has

$$\mathbb{E}^0\left[Z_{i,t}^{-1}|X_t = x\right] \leq 1 + \frac{C^{\text{te}}}{p_X^0(t,x)\sqrt{t}}\exp\left(-\frac{x^2}{2q'\sigma^{*2}t}\right).$$

Also, there exists a constant $K_q > 1$, such that,

$$\forall (t,x) \in]0,1]\times\mathbb{R} \;\; p_{i,X}(t,x) \geq \frac{1}{K_q\sqrt{t}}\exp\left(-\frac{2q'-1}{2q'\sigma^{*2}t}x^2\right), \;\; \forall q' > 1. \tag{3.25}$$

The final result is deduced from (3.24) and (3.25). □

3.7.2 Proofs of Section 3.3

We begin by providing the proof of Theorem 3.3.2 that relies in part on Proposition 3.3.1.

Proof of Theorem 3.3.2. From Proposition 3.3.1, we have the following inequality

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq 2\sum_{i=1}^{K}\mathbb{E}\left[|\widehat{\pi}_i(X) - \pi_i^*(X)|\right]. \tag{3.26}$$

We define $\bar{\mathbf{F}}$ the discretized version of $\mathbf{F}^*$,

$$\bar{\mathbf{F}} = (\bar{F}_1, \dots, \bar{F}_K), \quad \text{with} \quad \bar{F}_i(X) = \sum_{k=0}^{n-1} \left(\frac{b_i^*}{\sigma^{*2}}(X_{k\Delta}) \left(X_{(k+1)\Delta} - X_{k\Delta} \right) - \frac{\Delta}{2} \frac{b_i^{*2}}{\sigma^{*2}}(X_{k\Delta}) \right),$$

and for each $i \in \mathcal{Y}$, $\bar{\pi}_i^* = \phi_i\left(\bar{\mathbf{F}}\right)$ the discretized version of π_i^*, and $\bar{\pi}_i = \phi_i(\widehat{\boldsymbol{F}})$. From Equation (3.26), we deduce

$$\begin{aligned}
\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] &\leq 2 \left(\sum_{i=1}^{K} \mathbb{E}\left[|\widehat{\pi}_i(X) - \bar{\pi}_i(X)|\right] + \mathbb{E}\left[|\bar{\pi}_i(X) - \bar{\pi}_i^*(X)|\right] \right. \\
&\quad \left. + \sum_{i=1}^{K} \mathbb{E}\left[|\bar{\pi}_i^*(X) - \pi_i^*(X)|\right] \right) \\
&\leq 2 \sum_{i=1}^{K} \mathbb{E}\left[\left| \widehat{\phi}_i(\widehat{\boldsymbol{F}}(X)) - \phi_i(\widehat{\boldsymbol{F}}(X)) \right| \right] + 2 \sum_{i=1}^{K} \mathbb{E}\left[\left| \phi_i(\widehat{\boldsymbol{F}}(X)) - \phi_i(\bar{\mathbf{F}}(X)) \right| \right] \\
&\quad + 2 \sum_{i=1}^{K} \mathbb{E}\left[\left| \phi_i(\bar{\mathbf{F}}(X)) - \phi_i(\mathbf{F}^*(X)) \right| \right]. \qquad (3.27)
\end{aligned}$$

For the first term of the *r.h.s.* of the above inequality, we observe that for $(x_1, \dots, x_K) \in \mathbb{R}^K$, and $(i, j) \in \mathcal{Y}^2$ we have

$$\left| \frac{\partial}{\partial_{\mathfrak{p}_j^*}} \frac{\mathfrak{p}_i^* \exp(x_i)}{\sum_{k=1}^{K} \mathfrak{p}_k^* \exp(x_k)} \right| \leq \frac{1}{\mathfrak{p}_0^*}.$$

Therefore,

$$\sum_{i=1}^{K} \mathbb{E}\left[\left| \widehat{\phi}_i(\widehat{\boldsymbol{F}}(X)) - \phi_i(\widehat{\boldsymbol{F}}(X)) \right| \right] \leq C_{K,\mathfrak{p}_0^*} \sum_{k=1}^{K} \mathbb{E}\left[|\widehat{\mathfrak{p}}_k - \mathfrak{p}_k| \right] \leq \frac{C_{K,\mathfrak{p}_0^*}}{\sqrt{N}}. \qquad (3.28)$$

For the second term of Equation (3.27), since the softmax function is 1-Lipschitz, we have for $j \in \mathcal{Y}$

$$\mathbb{E}\left| \phi_j(\widehat{\boldsymbol{F}}(X)) - \phi_j(\bar{\mathbf{F}}(X)) \right| \leq \sum_{i=1}^{K} \mathbb{E}\left[\left| \widehat{F}_i(X) - \bar{F}_i(X) \right| \right].$$

We set $\xi(s) := k\Delta$, if $s \in [k\Delta, (k+1)\Delta)$, for $k \in [\![0, n-1]\!]$. We then deduce that

$$\begin{aligned}
\left| \widehat{F}_i(X) - \bar{F}_i(X) \right| \leq \int_0^1 \left| \left(\frac{\widehat{b}_i}{\widehat{\sigma}^2} - \frac{b_i^*}{\sigma^{*2}} \right) (X_{\xi(s)})\, b_Y^*(X_s) \right| ds + \frac{1}{2} \int_0^1 \left| \left(\frac{\widehat{b}_i^2}{\widehat{\sigma}^2} - \frac{b_i^{*2}}{\sigma^{*2}} \right) (X_{\xi(s)}) \right| ds \\
+ \left| \int_0^1 \left(\frac{\widehat{b}_i}{\widehat{\sigma}^2} - \frac{b_i^*}{\sigma^{*2}} \right) (X_{\xi(s)})\, \sigma^*(X_s) dW_s \right|,
\end{aligned}$$

which implies

$$\begin{aligned}
\mathbb{E}\left[\left| \widehat{F}_i(X) - \bar{F}_i(X) \right| \right] &\leq \mathbb{E}\left[\int_0^1 \left| \left(\frac{\widehat{b}_i}{\widehat{\sigma}^2} - \frac{b_i^*}{\sigma^{*2}} \right) (X_{\xi(s)})\, b_Y^*(X_s) \right| ds \right] + \frac{1}{2} \mathbb{E}\left[\int_0^1 \left| \left(\frac{\widehat{b}_i^2}{\widehat{\sigma}^2} - \frac{b_i^{*2}}{\sigma^{*2}} \right) (X_{\xi(s)}) \right| ds \right] \\
&\quad + \mathbb{E}\left[\int_0^1 \left(\frac{\widehat{b}_i}{\widehat{\sigma}^2} - \frac{b_i^*}{\sigma^{*2}} \right)^2 (X_{\xi(s)})\, \sigma^{*2}(X_s) ds \right].
\end{aligned}$$

Since for all x, $\sigma^*(x) \geq \sigma_0^*$, and $\widehat{\sigma} \geq \sigma_0$, we get

$$\begin{cases}
\left| \frac{\widehat{b}_i}{\widehat{\sigma}^2}(x) - \frac{b_i^*}{\sigma^{*2}}(x) \right| \leq \sigma_0^{-2} \left| \widehat{b}_i(x) - b_i^*(x) \right| + \sigma_0^{-2} \sigma_0^{*-2} \left| b_i^*(x) \right| \left| \widehat{\sigma}^2(x) - \sigma^{*2}(x) \right|, \\
\left| \frac{\widehat{b}_i^2}{\widehat{\sigma}^2}(x) - \frac{b_i^{*2}}{\sigma^{*2}}(x) \right| \leq \sigma_0^{-2} \left| \widehat{b}_i(x) + b_i^*(x) \right| \left| \widehat{b}_i(x) - b_i^*(x) \right| + \sigma_0^{-2} \sigma_0^{*-2} \left| b_i^*(x) \right|^2 \left| \widehat{\sigma}^2(x) - \sigma^{*2}(x) \right|.
\end{cases} \qquad (3.29)$$

Hence, as $\widehat{b}_i(x) \leq b_{\max}$, and $\mathbb{E}\left[\sup_{t\in[0,1]} |b_i^*(X_t)|\right] \leq C_1$, the above inequalities and the Cauchy-Schwarz inequality yield

$$\mathbb{E}\left|\widehat{F}_i(X) - \bar{F}_i(X)\right| \leq C_{\sigma_0^*}\sigma_0^{-2}\left(b_{\max}\mathbb{E}\left\|\widehat{b}_i - b_i^*\right\|_n + \mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^2\right\|_n\right).$$

Therefore, we have,

$$\sum_{i=1}^{K}\mathbb{E}\left[\left|\phi_i(\widehat{\boldsymbol{F}}(X)) - \phi_i(\bar{\mathbf{F}}(X))\right|\right] \leq C_{K,\sigma_0^*}\sigma_0^{-2}\sum_{i=1}^{K}\left(b_{\max}\mathbb{E}\left\|\widehat{b}_i - b_i^*\right\|_n + \mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^2\right\|_n\right). \tag{3.30}$$

Finally, the last term is bounded as follows. We first observe that for all $i \in \mathcal{Y}$

$$\begin{aligned}\mathbb{E}\left[\left|\bar{F}_i(X) - F_i^*(X)\right|^2\right] \leq\ & 3\mathbb{E}\int_0^1\left(\frac{b_i^*\left(X_{\xi(s)}\right)}{\sigma^{*2}\left(X_{\xi(s)}\right)} - \frac{b_i^*\left(X_s\right)}{\sigma^{*2}(X_s)}\right)^2 b_Y^{*2}(X_s)ds \\ & + 3\mathbb{E}\int_0^1\left(\frac{b_i^{*2}\left(X_{\xi(s)}\right)}{\sigma^{*2}\left(X_{\xi(s)}\right)} - \frac{b_i^{*2}\left(X_s\right)}{\sigma^{*2}(X_s)}\right)^2 ds + 3\mathbb{E}\int_0^1\left(\frac{b_i^*\left(X_{\xi(s)}\right)}{\sigma^{*2}\left(X_{\xi(s)}\right)} - \frac{b_i^*\left(X_s\right)}{\sigma^{*2}(X_s)}\right)^2 \sigma^{*2}(X_s)ds.\end{aligned}$$

Using again that $\sigma^*(\cdot) \geq \sigma_0^*$, and $\mathbb{E}\left[\sup_{t\in[0,1]} |b_i^*(X_t)|^q\right] \leq C$ for $q \geq 1$ (by Assumption 3.2.1), the Cauchy-Schwarz inequality implies

$$\begin{aligned}\mathbb{E}\left[\left|\bar{F}_i(X) - F_i^*(X)\right|^2\right] \leq C_{\sigma_0^*}\Bigg(&\int_0^1\mathbb{E}\left[\left|b_i^*\left(X_{\xi(s)}\right) - b_i^*(X_s)\right|^2\right]ds \\ &+ \int_0^1\sqrt{\mathbb{E}\left[\left|b_i^*\left(X_{\xi(s)}\right) - b_i^*(X_s)\right|^4\right]}ds + \int_0^1\sqrt{\mathbb{E}\left[\left|\sigma^{*2}\left(X_{\xi(s)}\right) - \sigma^{*2}(X_s)\right|^4\right]}ds\Bigg).\end{aligned}$$

Finally, since the functions b_i^*, and σ^* are Lipschitz, we deduce from Lemma 3.7.1 that

$$\mathbb{E}\left[\left|\bar{F}_i(X) - F_i^*(X)\right|^2\right] \leq C_{\sigma_0^*}\Delta,$$

which implies together with the fact that the sofmax function is 1-Lipschitz and the Jensen inequality that

$$\sum_{i=1}^{K}\mathbb{E}\left[\left|\phi_i(\bar{\mathbf{F}}(X)) - \phi_i(\mathbf{F}^*(X))\right|\right] \leq C_{K,\sigma_0^*}\sqrt{\Delta}. \tag{3.31}$$

In view of Equation 3.27, the combination of Equations (3.28), and (3.30), and (3.31) yields the desired result. □

Proof of Proposition 3.3.3. We consider h a L-Lipschitz function. We define the spline-approximation $\tilde{h}$ of h by

$$\tilde{h}(x) := \sum_{\ell=-M}^{K^*-1} h(u_\ell)B_\ell(x), \quad \forall x \in \mathbb{R}.$$

First, we note that $\tilde{h} \in \mathcal{S}_{K^*,M}$. Indeed, since h is L-Lipschitz, there exists $C_L > 0$ such that for A large enough,

$$|h(x)| \leq C_L(1 + |x|) \leq CA, \quad \forall x \in [-A, A].$$

Therefore, for N large enough, we have

$$|h(x)| \leq A\log^{1/2}(N).$$

Then, we deduce

$$\sum_{\ell=-M}^{\Xi-1} h^2(u_\ell) \leq (\Xi + M)A^2\log(N).$$

For $x \in [-A, A]$, there exists $0 \leq \ell_0 \leq \Xi - 1$ such that $x \in [u_{\ell_0}, u_{\ell_0+1})$. We use the following property of the B-spline basis

$$B_\ell(x) = 0, \quad \text{if} \quad x \notin [u_\ell, u_{\ell+M+1}), \quad \ell = -M, \ldots, K_N + M.$$

Hence, for $x \in [u_{\ell_0}, u_{\ell_0+1})$, we have $B_\ell(x) = 0$ for $\ell \leq \ell_0 - M - 1$, and $\ell \geq \ell_0 + M$. Thus,

$$\begin{aligned}
\left|\tilde{h}(x) - h(x)\right| &\leq \sum_{\ell=-M}^{\Xi-1} |h(u_\ell) - h(x)|\, B_\ell(x) \\
&= \sum_{\ell=\ell_0-M}^{\ell_0} |h(u_\ell) - h(x)|\, B_l(x) \\
&\leq \max_{\ell=\ell_0-M,\ldots,\ell_0} |h(u_\ell) - h(x)| \\
&\leq L(u_{\ell_0+1} - u_{\ell_0-M}) \leq \frac{2L(M+1)A}{\Xi},
\end{aligned}$$

which concludes the proof. □

Proof of Theorem 3.3.4. The proof is divided in two parts. The first part establishes the rates of convergence of the drift estimators, and the second part is devoted to the study of the rates of convergence of the diffusion coefficient estimator.

Rates of convergence for drift estimator. Let $i \in \{1, \ldots, K\}$. We introduce, on the random event $\{N_i > 1\}$, and with $A_{N_i} = \log(N_i)$, the function,

$$\bar{b}_i := b_i^* \mathbb{1}_{(-A_{N_i}, A_{N_i})}.$$

We recall that $N_i = \sum_{j=1}^N \mathbb{1}_{\{Y_j = i\}}$ is the random number of paths in the class number i. For a function h, we introduce the empirical norm of class i on the event $\{N_i > 1\}$ as

$$\|h\|_{n,N_i}^2 := \frac{1}{nN_i} \sum_{j \in \mathcal{I}_i} \sum_{k=0}^{n-1} h^2(X_{k\Delta}^j).$$

We first observe that

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_n\right] = \mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_n \mathbb{1}_{\{N_i > 1\}}\right] + \mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_n \mathbb{1}_{\{N_i \leq 1\}}\right].$$

Let us work at first on the event $\{N_i > 1\}$. For all $i \in \mathcal{Y}$, we define the following conditional expectation

$$\mathbb{E}_i[.] = \mathbb{E}[.|\mathbb{1}_{\{Y_1=i\}}, \ldots, \mathbb{1}_{\{Y_N=i\}}].$$

We apply Proposition 3.3.3, and Proposition 3.2 of [30] on the event $\{N_i > 1\}$ and deduce that

$$\mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,N_i}^2\right] \leq C\left(\frac{A_{N_i}^2}{K_{N_i}^2} + \sqrt{\frac{K_{N_i} A_{N_i}^2 \log(N)}{N_i}} + \Delta\right). \tag{3.32}$$

Now, for all $i \in \mathcal{Y}$, let us write

$$\mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,i}^2\right] = \mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,i}^2\right] - 2\mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,N_i}^2\right] + 2\mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,N_i}^2\right]. \tag{3.33}$$

For $h \in \mathcal{S}_{K_{N_i},M}$, we denote by $\bar{h}$ its thresholded counterpart

$$\bar{h}(\cdot) := h(\cdot)\mathbb{1}_{\{|h(\cdot)| \leq A_{N_i} \log^{1/2}(N)\}} + \operatorname{sgn}(h(\cdot)) A_{N_i} \log^{1/2}(N) \mathbb{1}_{\{|h(\cdot)| > A_{N_i} \log^{1/2}(N)\}}.$$

We also denote $\mathcal{H}_{K_{N_i},M} := \{\bar{h},\ h \in \mathcal{S}_{K_{N_i},M}\}$. Then, on the event $\{N_i > 1\}$, we have that

$$\begin{aligned}\mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,i}^2\right] - 2\mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,N_i}^2\right] &\leq \mathbb{E}_i\left[\sup_{\bar{h}\in\mathcal{H}_{K_{N_i},M}} \left\|\bar{h} - \bar{b}_i\right\|_{n,i}^2 - 2\left\|\bar{h} - \bar{b}_i\right\|_{n,N_i}^2\right] \\ &\leq \mathbb{E}_i\left[\sup_{g\in\mathcal{G}_{K_{N_i},M}} \mathbb{E}_{X|Y=i}\left[g(\bar{X}) - \frac{2}{N_i}\sum_{j\in\mathcal{I}_i} g(\bar{X}^j)\right]\right],\end{aligned}$$

with $\mathcal{G}_{K_{N_i},M} = \{(x_1,\dots,x_n) \mapsto \frac{1}{n}\sum_{k=1}^n \left|\bar{h}(x_k) - \bar{b}_i(x_k)\right|^2, \bar{h} \in \mathcal{H}_{K_{N_i},M}\}$. For each $g \in \mathcal{G}_{K_{N_i},M}$ and $x \in \mathbb{R}$, we have on the event $\{N_i > 1\}$,

$$0 \leq g(x) \leq 4A_{N_i}^2 \log(N).$$

Furthermore, we have that (see [30])

$$\mathcal{N}_\infty\left(\varepsilon, \mathcal{G}_{K_{N_i},M}\right) \leq \left(\frac{12(K_{N_i}+M)A_{N_i}^2\log(N)}{\varepsilon}\right)^{K_{N_i}+M} \leq \left(\frac{12(K_N+M)\log^3(N)}{\varepsilon}\right)^{K_N+M}.$$

Therefore, we deduce from Lemma A.2 in [30] with $\varepsilon = \dfrac{12(K_N+M)\log^3(N)}{N_i}$, Equation (3.32), and Equation (3.33), that on the event $\{N_i > 1\}$ with $A_{N_i} = \log(N_i)$

$$\mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,i}^2\right] \leq C\left(\frac{\log^2(N_i)}{K_{N_i}^2} + \sqrt{\frac{K_{N_i}\log^2(N_i)\log(N)}{N_i}} + \frac{\log^2(N_i)\log^2(N)K_{N_i}}{N_i} + \Delta\right). \tag{3.34}$$

Thus, choosing $K_{N_i} \propto (N_i\log(N_i))^{1/5}$ and for $\log(N_i) \leq \log(N)$, we obtain from Equation (3.34) that

$$\mathbb{E}_i\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,i}^2\right] \leq C\left(\frac{\log^{8/5}(N)}{N_i^{2/5}} + \frac{\log^{21/5}(N)}{N_i^{4/5}} + \Delta\right). \tag{3.35}$$

Using Jensen's inequality, we have

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,i}\mathbb{1}_{\{N_i>1\}}\right] \leq \sqrt{\mathbb{E}\left[\left\|\widehat{b}_i - \bar{b}_i\right\|_{n,i}^2\mathbb{1}_{\{N_i>1\}}\right] + \mathbb{E}\left[\left\|\bar{b}_i - b_i^*\right\|_{n,i}^2\mathbb{1}_{\{N_i>1\}}\right]}.$$

Finally, let us study then the error $\left\|\bar{b}_i - b_i^*\right\|_{n,i}^2$. On the event $\{N_i > 1\}$, we observe with the Cauchy-Schwarz inequality

$$\begin{aligned}\left\|\bar{b}_i - b_i^*\right\|_{n,i}^2 &= \mathbb{E}_{X|Y=i}\left[\frac{1}{n}\sum_{k=1}^n \left(b_i^*(X_{k\Delta})\right)^2 \mathbb{1}_{\{|X_{k\Delta}|>A_{N_i}\}} |\, \mathbb{1}_{Y_1=i},\dots,\mathbb{1}_{Y_N=i}\right] \\ &\leq C\sqrt{\sup_{t\in[0,1]} \mathbb{P}_{X|Y=i}\left(|X_t| \geq A_{N_i}|\, \mathbb{1}_{Y_1=i},\dots,\mathbb{1}_{Y_N=i}\right)},\end{aligned}$$

since $\sup_{t\in[0,1]}\mathbb{E}\left[b_i^*(X_t)^4\right] \leq C$. For $A_{N_i} = \log(N_i)$ and from Lemma 3.7.3, we obtain on the event $\{N_i > 1\}$

$$\left\|\bar{b}_i - b_i^*\right\|_{n,i}^2 \leq C\exp\left(-\frac{C_2}{2}\log^2(N_i)\right),$$

which, for N_i a.s. large enough yields

$$\left\|\bar{b}_i - b_i^*\right\|_{n,i} \leq CN_i^{-1/2}.$$

This result leads us to obtain, from Equation (3.35), that

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,i}\mathbb{1}_{\{N_i>1\}}\right] \leq C\left(\mathbb{E}\left[\left(\frac{\log^{4/5}(N)}{N_i^{1/5}} + \frac{\log^{21/10}(N)}{N_i^{2/5}}\right)\mathbb{1}_{\{N_i>1\}}\right] + \sqrt{\Delta}\right).$$

Using Jensen's inequality, we obtain

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,i} \mathbb{1}_{\{N_i>1\}}\right] \leq C\left(\log^{4/5}(N)\left(\mathbb{E}\left[\frac{\mathbb{1}_{N_i>1}}{N_i}\right]\right)^{1/5} + \log^{21/10}(N)\left(\mathbb{E}\left[\frac{\mathbb{1}_{N_i>1}}{N_i}\right]\right)^{2/5} + \sqrt{\Delta}\right).$$

To finish the proof, since for all $i \in \mathcal{Y}$, $N_i \sim \mathcal{B}(N, \mathfrak{p}_i^*)$ we use Lemma 4.1 in [54] to deduce that

$$\mathbb{E}\left[\frac{\mathbb{1}_{N_i>1}}{N_i}\right] \leq \mathbb{E}\left[\frac{\mathbb{1}_{N_i>0}}{N_i}\right] \leq \frac{2}{\mathfrak{p}_i^* N} \leq \frac{2}{\mathfrak{p}_0^* N},$$

and finally, there exists a constant $C > 0$ such that

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,i} \mathbb{1}_{\{N_i>1\}}\right] \leq C\left(\left(\frac{\log^4(N)}{N\mathfrak{p}_0^*}\right)^{1/5} + \sqrt{\Delta}\right). \tag{3.36}$$

To conclude the proof for the rates of convergence of the drift coefficient, we observe that since $\widehat{b}_i$ is bounded by $\log^{3/2}(N)$ and $\sup_{t\in[0,1]} \mathbb{E}\left[b_i^*(X_t)^2\right] < +\infty$, we have for N large enough,

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,i} \mathbb{1}_{\{N_i\leq 1\}}\right] \leq 2\log^{3/2}(N)\left(\mathbb{P}\left(N_i = 0\right) + \mathbb{P}\left(N_i = 1\right)\right). \tag{3.37}$$

Since N_i is distributed according to a Binomial distribution with parameters $(N, \mathfrak{p}_i^*)$. We deduce that

$$\mathbb{P}\left(N_i = 0\right) = \exp\left(N\log(1-\mathfrak{p}_i^*)\right), \ \ \mathbb{P}\left(N_i = 1\right) = \frac{\mathfrak{p}_i^*}{1-\mathfrak{p}_i^*}\exp\left(N\log(1-\mathfrak{p}_i^*)\right). \tag{3.38}$$

Hence, gathering Equation (3.36), Equation (3.37) and Equation (3.38), and choosing $\Delta = O(1/N)$, we obtain for each label $i \in \mathcal{Y}$,

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,i}\right] = \mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,i} \mathbb{1}_{N_i>1}\right] + \mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,i} \mathbb{1}_{N_i\leq 1}\right] = \mathrm{O}\left(\left(\frac{\log^4(N)}{N\mathfrak{p}_0^*}\right)^{1/5}\right).$$

Diffusion coefficient: rates of convergence. We estimate the square σ^{*2} of the diffusion coefficient as solution of the following regression model

$$\frac{(X^j_{(k+1)\Delta} - X^j_{k\Delta})^2}{\Delta} = \sigma^{*2}(X^j_{k\Delta}) + \zeta^j_{k\Delta} + R^j_{k\Delta}, \tag{3.39}$$

where $\zeta^j_{k\Delta} := \zeta^{j,1}_{k\Delta} + \zeta^{j,2}_{k\Delta} + \zeta^{j,3}_{k\Delta}$ with

$$\zeta^{j,1}_{k\Delta} := \frac{1}{\Delta}\left[\left(\int_{k\Delta}^{(k+1)\Delta} \sigma^*\left(X^j_s\right) dW^j_s\right)^2 - \int_{k\Delta}^{(k+1)\Delta} \sigma^{*2}\left(X^j_s\right) ds\right],$$

$$\zeta^{j,2}_{k\Delta} := \frac{2}{\Delta}\int_{k\Delta}^{(k+1)\Delta} ((k+1)\Delta - s)\sigma^{*\prime}\left(X^j_s\right)\sigma^{*2}\left(X^j_s\right) dW^j_s,$$

$$\zeta^{j,3}_{k\Delta} := 2b_Y^*\left(X^j_{k\Delta}\right)\int_{k\Delta}^{(k+1)\Delta} \sigma^*\left(X^j_s\right) dW^j_s,$$

and $R^j_{k\Delta} := R^{j,1}_{k\Delta} + R^{j,2}_{k\Delta} + R^{j,3}_{k\Delta}$ with,

$$R^{j,1}_{k\Delta} := \frac{1}{\Delta}\left(\int_{k\Delta}^{(k+1)\Delta} b_Y^*\left(X^j_s\right) ds\right)^2, \quad R^{j,2}_{k\Delta} := \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta} ((k+1)\Delta - s)\phi_Y\left(X^j_s\right) ds, \tag{3.40}$$

$$R^{j,3}_{k\Delta} := \frac{2}{\Delta}\left(\int_{k\Delta}^{(k+1)\Delta}\left(b_Y^*\left(X^j_s\right) - b_Y^*\left(X_{k\Delta}\right)\right) ds\right)\left(\int_{k\Delta}^{(k+1)\Delta} \sigma^*\left(X^j_s\right) dW^j_s\right), \tag{3.41}$$

where $\phi_Y := b_Y^*\sigma^{*\prime}\sigma^* + \left[\sigma^{*\prime\prime}\sigma^* + (\sigma^{*\prime})^2\right]\sigma^{*2}$. We prove in the sequel that $\zeta^{j,1}_{k\Delta}$ is the error term, and all the other terms are negligible residuals. We remind the reader that the estimator $\widehat{\sigma}^2$ of σ^{*2} is given in (3.10). We rely on the following result:

Lemma 3.7.6. *Under Assumption 3.2.1 and for $\tilde{A}_N = \log(N)$, the following holds*

$$\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_{n,N}^2 \leq 3 \inf_{h \in \mathcal{S}_{\tilde{K}_N, M}} \|h - \sigma^{*2}\|_n^2 + C\left(\sqrt{\frac{\tilde{K}_N \log^3(N)}{Nn}} + \Delta_n^2\right),$$

where $C > 0$ is a constant depending on σ_1, and where

$$\|h\|_{n,N}^2 = \frac{1}{nN}\sum_{j=1}^{N}\sum_{k=0}^{n-1} h^2(X_{k\Delta}^j).$$

The empirical error of the estimator $\widehat{\sigma}^2$ is given by

$$\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_n^2 = 2\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_{n,N}^2 + \left[\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_n^2 - 2\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_{n,N}^2\right].$$

Since $\tilde{A}_N = \log(N)$, let us define $\mathcal{H}^\sigma$ as the set of functions $\bar{h}$ such that there exists a function $h \in \mathcal{S}_{K_N, M}$ satisfying

$$\bar{h} = h(x)\mathbb{1}_{\{\frac{1}{\log(N)} \leq h(x) \leq \log^{3/2}(N)\}} + \log^{3/2}(N)\mathbb{1}_{h(x) > \log^{3/2}(N)\}} + \frac{1}{\log(N)}\mathbb{1}_{\{h(x) \leq \frac{1}{\log(N)}\}}.$$

Using then an $\varepsilon-$net $\mathcal{H}^{\sigma,\varepsilon}$ of $\mathcal{H}^\sigma$ with $\varepsilon = \frac{12(\tilde{K}_N + M)\log^3(N)}{N}$, we finally obtain (see [30], Lemma A.2)

$$\begin{aligned}\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_n^2 - 2\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_{n,N}^2 &\leq \mathbb{E}\left[\sup_{\bar{h} \in \mathcal{H}^\sigma}\left\{\mathbb{E}\left\|\bar{h} - \sigma^{*2}\right\|_n^2 - 2\mathbb{E}\left\|\bar{h} - \sigma^{*2}\right\|_{n,N}^2\right\}\right] \\ &\leq C\frac{\tilde{K}_N \log^4(N)}{N}.\end{aligned}$$

Thus, as $\Delta_n = \mathrm{O}(1/N)$,

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_n^2\right] &\leq 3 \inf_{h \in \mathcal{S}_{\tilde{K}_N, M}} \|h - \sigma^{*2}\|_n^2 + C\left(\frac{\sqrt{\tilde{K}_N \log^3(N)}}{N} + \frac{\tilde{K}_N \log^4(N)}{N} + \frac{1}{N^2}\right) \\ &\leq 3 \inf_{h \in \mathcal{S}_{\tilde{K}_N M}} \|h - \sigma^{*2}\|_n^2 + C\frac{\tilde{K}_N \log^4(N)}{N},\end{aligned}$$

for N large enough. According to Proposition 3.3.3, the bias term satisfies

$$\inf_{h \in \mathcal{S}_{\tilde{K}_N, M}} \|h - \sigma^{*2}\|_n^2 \leq C\frac{\log^2(N)}{\tilde{K}_N^2}.$$

Taking $\tilde{K}_N = (N\log(N))^{1/5}$ leads to

$$\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_n \leq C_2\left(\frac{\log^4(N)}{N}\right)^{1/5}.$$

This concludes the proof of Theorem 3.3.4. □

Proof of Lemma 3.7.6 . Denote by

$$\gamma_{N,n}(h) = \frac{1}{nN}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\left(U_{k\Delta}^j - h(X_{k\Delta}^j)\right)^2,$$

the least square contrast appearing in (3.9). For all $h \in \mathcal{S}_{K_N, M}$, we deduce that

$$\gamma_{n,N}(\widehat{\sigma}^2) - \gamma_{n,N}(\sigma^{*2}) \leq \gamma_{n,N}(h) - \gamma_{n,N}(\sigma^{*2}). \tag{3.42}$$

Using (3.39), we have for all $h \in \mathcal{S}_{\tilde{K}_N,M}$,

$$\gamma_{n,N}(h)-\gamma_{n,N}(\sigma^{*2}) = \left\|h-\sigma^{*2}\right\|_{n,N}^2 + 2\nu_1(\sigma^{*2}-h)+2\nu_2(\sigma^{*2}-h)+2\nu_3(\sigma^{*2}-h)+2\mu(\sigma^{*2}-h), \quad (3.43)$$

where

$$\nu_i(h) = \frac{1}{nN}\sum_{j=1}^{N}\sum_{k=0}^{n-1} h(X_{k\Delta}^j)\zeta_{k\Delta}^{j,i}, \quad i\in\{1,2,3\}, \quad \mu(h) = \frac{1}{nN}\sum_{j=1}^{N}\sum_{k=0}^{n-1} h(X_{k\Delta}^j)R_{k\Delta}^j, \quad (3.44)$$

we derive from Equations (3.42) and (3.43) that for all $h \in \mathcal{S}_{\tilde{K}_N,M}$,

$$\mathbb{E}\left\|\widehat{\sigma}^2-\sigma^{*2}\right\|_{n,N}^2 \leq \inf_{h\in\mathcal{S}_{\tilde{K}_N,M}}\|h-\sigma^{*2}\|_n^2 + 2\sum_{i=1}^{3}\mathbb{E}\left[\nu_i(\widehat{\sigma}^2-h)\right] + 2\mathbb{E}\left[\mu(\widehat{\sigma}^2-h)\right]. \quad (3.45)$$

For all $i\in\{1,2,3\}$ and for all $h\in\mathcal{S}_{\tilde{K}_N,M}$, taking the constraints (3.6) into account, one has

$$\mathbb{E}\left[\nu_i\left(\widehat{\sigma}^2-h\right)\right] \leq \sqrt{2(\tilde{K}_N+M)\log^3(N)}\sqrt{\sum_{\ell=-M}^{\tilde{K}_N-1}\mathbb{E}\left[\nu_i^2(B_{\ell,M,\mathbf{u}})\right]}. \quad (3.46)$$

1. Upper bound of $\sum_{\ell=-M}^{\tilde{K}_N-1}\mathbb{E}\left[\nu_1^2(B_{\ell,M,\mathbf{u}})\right]$. According to Equation (3.44), we have

$$\forall \ell\in[\![-M,\tilde{K}_N-1]\!],\ \nu_1(B_{\ell,M,\mathbf{u}}) = \frac{1}{nN}\sum_{j=1}^{N}\sum_{k=0}^{n-1} B_{\ell,M,\mathbf{u}}(X_{k\Delta}^j)\zeta_{k\Delta}^{j,1},$$

where $\zeta_{k\Delta}^{j,1} = \frac{1}{\Delta}\left[\left(\int_{k\Delta}^{(k+1)\Delta}\sigma^*(X_s^j)dW_s^j\right)^2 - \int_{k\Delta}^{(k+1)\Delta}\sigma^{*2}(X_s^j)ds\right]$ is a martingale satisfying

$$\mathbb{E}\left[\zeta_{k\Delta}^{1,1}|\mathcal{F}_{k\Delta}^1\right]=0 \text{ and } \mathbb{E}\left[\left(\zeta_{k\Delta}^{1,1}\right)^2|\mathcal{F}_{k\Delta}^1\right] \leq \frac{1}{\Delta^2}\mathbb{E}\left[\left(\int_{k\Delta}^{(k+1)\Delta}\sigma^{*2}(X_s^1)ds\right)^2\right] \leq C\sigma_1^{*4},$$

with $(\mathcal{F}_t^1)_{t\geq 0}$ the natural filtration associated with the Brownian motion W^1. We derive that

$$\begin{aligned}\sum_{\ell=-M}^{\tilde{K}_N-1}\mathbb{E}\left[\nu_1^2(B_{\ell,M,\mathbf{u}})\right] =& \frac{1}{Nn^2}\sum_{\ell=-M}^{\tilde{K}_N-1}\mathbb{E}\left[\left(\sum_{k=0}^{n-1}B_{\ell,M,\mathbf{u}}(X_{k\Delta}^j)\zeta_{k\Delta}^{1,1}\right)^2\right]\\ =& \frac{1}{Nn^2}\mathbb{E}\left[\sum_{k=0}^{n-1}\sum_{\ell=-M}^{\tilde{K}_N-1}B_{\ell,M,\mathbf{u}}^2(X_{k\Delta}^1)\left(\zeta_{k\Delta}^{1,1}\right)^2\right]\\ \leq& \frac{C}{Nn},\end{aligned}$$

where C is a constant depending on σ^*, for each $k\in[\![0,n-1]\!]$, $\sum_{\ell=-M}^{\tilde{K}_N-1}B_{\ell,M,\mathbf{u}}^2(X_{k\Delta}^1)\leq 1$ since $\sum_{\ell=-M}^{\tilde{K}_N-1}B_{\ell,M,\mathbf{u}}(X_{k\Delta}^1)=1$ and $B_{\ell,M,\mathbf{u}}(X_{k\Delta}^1)\leq 1$ for all $\ell=-M,\cdots,\tilde{K}_N-1$.

2. Upper bound of $\sum_{\ell=-M}^{\tilde{K}_N-1}\mathbb{E}\left[\nu_2^2(B_{\ell,M,\mathbf{u}})\right]$. For all $k\in[\![0,n-1]\!]$ and for all $s\in[0,1]$, set $\xi(s)=k\Delta$ if $s\in[k\Delta,(k+1)\Delta)$. We have:

$$\begin{aligned}&\sum_{\ell=-M}^{\tilde{K}_N-1}\mathbb{E}\left[\nu_2^2(B_{\ell,M,\mathbf{u}})\right]\\ &\quad= \frac{4}{Nn^2}\sum_{\ell=-M}^{\tilde{K}_N-1}\mathbb{E}\left[\left(\sum_{k=0}^{n-1}\int_{k\Delta}^{(k+1)\Delta}B_{\ell,M,\mathbf{u}}(X_{k\Delta}^1)((k+1)\Delta-s)\sigma^{*\prime}(X_s^1)\sigma^{*2}(X_s^1)dW_s\right)^2\right]\\ &\quad= \frac{4}{Nn^2}\sum_{\ell=-M}^{\tilde{K}_N-1}\mathbb{E}\left[\left(\int_0^1 B_{\ell,M,\mathbf{u}}(X_{\xi(s)}^1)(\xi(s)+\Delta-s)\sigma^{*\prime}(X_s^1)\sigma^{*2}(X_s^1)dW_s\right)^2\right]\\ &\quad\leq \frac{C}{Nn^2},\end{aligned}$$

where the constant $C > 0$ depends on the diffusion coefficient.

3. Upper bound of $\sum_{\ell=-M}^{\tilde{K}_N-1} \mathbb{E}\left[\nu_3^2(B_{\ell,M,\mathbf{u}})\right]$. We have:

$$\begin{aligned}\sum_{\ell=-M}^{\tilde{K}_N-1} \mathbb{E}\left[\nu_3^2(B_{\ell,M,\mathbf{u}})\right] &= \frac{4}{Nn^2}\sum_{\ell=-M}^{\tilde{K}_N-1} \mathbb{E}\left[\left(\sum_{k=0}^{n-1}\int_{k\Delta}^{(k+1)\Delta} B_{\ell,M,\mathbf{u}}(X^1_{k\Delta}) b^*_Y(X^1_{k\Delta})\sigma^*(X^1_s)dW_s\right)^2\right] \\ &= \frac{4}{Nn^2}\sum_{\ell=-M}^{\tilde{K}_N-1} \mathbb{E}\left[\left(\int_0^1 B_{\ell,M,\mathbf{u}}(X^1_{\eta(s)}) b^*_Y(X^1_{\eta(s)})\sigma^*(X^1_s)dW_s\right)^2\right] \\ &\leq \frac{4}{Nn^2}\mathbb{E}\left[\int_0^1 \sum_{\ell=-M}^{\tilde{K}_N-1} B^2_{\ell,M,\mathbf{u}}(X^1_{\eta(s)}) b^{*2}_Y(X^1_{\eta(s)})\sigma^{*2}(X^1_s)ds\right].\end{aligned}$$

Since for all $x \in \mathbb{R}$, $b^{*2}_Y(x) \leq C_0(1+x^2)$, $\sigma^{*2}(x) \leq \sigma_1^{*2}$ and $\sup_{t\in[0,1]} \mathbb{E}\left(|X_t|^2\right) < \infty$, there exists a constant $C > 0$ depending on the upper bound σ_1^* of the diffusion coefficient such that

$$\sum_{\ell=-M}^{\tilde{K}_N-1} \mathbb{E}\left[\nu_3^2(B_{\ell,M,\mathbf{u}})\right] \leq \frac{C}{Nn^2}.$$

We finally deduce from Equations (2.62) and (3.46) that for all $h \in \mathcal{S}_{\tilde{K}_N,M}$,

$$\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|^2_{n,N} \leq \inf_{h\in\mathcal{S}_{\tilde{K}_N,M}} \|h - \sigma^{*2}\|^2_n + C\sqrt{\frac{(\tilde{K}_N + M)\log^3(N)}{Nn}} + 2\mathbb{E}\left[\mu(\widehat{\sigma}^2 - h)\right]. \tag{3.47}$$

It remains to obtain an upper bound of the term $\mu(\widehat{\sigma}^2 - h)$. Notice that for $a > 0$, x and $y \in \mathbb{R}$,

$$2xy = 2\frac{x}{\sqrt{a}} \times \sqrt{a}y \leq \frac{x^2}{a} + ay^2.$$

Then, for all $h \in \mathcal{S}_{\tilde{K}_N,M}$ and $a > 0$,

$$2\mu\left(\widehat{\sigma}^2 - h\right) \leq \frac{2}{a}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|^2_{n,N} + \frac{2}{a}\left\|h - \sigma^{*2}\right\|^2_{n,N} + \frac{a}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\left(R^j_{k\Delta}\right)^2.$$

We set $a = 4$ and from Equation (3.47) we deduce that,

$$\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|^2_{n,N} \leq 3\inf_{h\in\mathcal{S}_{\tilde{K}_N,M}} \|h - \sigma^{*2}\|^2_n + C\sqrt{\frac{(\tilde{K}_N + M)\log^3(N)}{Nn}} + \frac{4}{Nn}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\mathbb{E}\left[\left(R^j_{k\Delta}\right)^2\right]. \tag{3.48}$$

We have

$$\mathbb{E}\left[\left(R^j_{k\Delta}\right)^2\right] \leq 3\left(\mathbb{E}\left[\left(R^{j,1}_{k\Delta}\right)^2\right] + \mathbb{E}\left[\left(R^{j,2}_{k\Delta}\right)^2\right] + \mathbb{E}\left[\left(R^{j,3}_{k\Delta}\right)^2\right]\right),$$

where for all $j \in [\![1, N]\!]$ and $k \in [\![0, n-1]\!]$, $R^{j,1}_{k\Delta}$, $R^{j,2}_{k\Delta}$ and $R^{j,3}_{k\Delta}$ are given in Equations (3.40) and (3.41). There exist constants $C_1, C_2, C_3 > 0$ such that

$$\begin{aligned}&\mathbb{E}\left[\left(R^{j,1}_{k\Delta}\right)^2\right] \leq \mathbb{E}\left[\left(\int_{k\Delta}^{(k+1)\Delta} b^{*2}_Y\left(X^j_{k\Delta}\right)ds\right)^2\right] \leq \Delta\mathbb{E}\left[\int_{k\Delta}^{(k+1)\Delta} b^{*4}_Y\left(X^j_{k\Delta}\right)ds\right] \leq C_1\Delta^2 \\ &\mathbb{E}\left[\left(R^{j,2}_{k\Delta}\right)^2\right] \leq \frac{1}{\Delta^2}\int_{k\Delta}^{(k+1)\Delta}((k+1)\Delta - s)^2 ds \int_{k\Delta}^{(k+1)\Delta}\mathbb{E}\left[\phi^2_Y\left(X^j_s\right)\right]ds \leq C_2\Delta^2 \\ &\mathbb{E}\left[\left(R^{j,3}_{k\Delta}\right)^2\right] \leq \frac{4}{\Delta^2}\mathbb{E}\left[\Delta\int_{k\Delta}^{(k+1)\Delta} L_0^2\left|X^j_s - X^j_{k\Delta}\right|^2 ds\left(\int_{k\Delta}^{(k+1)\Delta}\sigma^*(X^j_s)dW_s\right)^2\right] \leq C_3\Delta^2.\end{aligned}$$

We deduce from Equation (3.48) that there exists a constant $C > 0$ depending on σ_1^* and M such that,

$$\mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_{n,N}^2 \leq 3 \inf_{h \in \mathcal{S}_{\tilde{K}_N, M}} \|h - \sigma^{*2}\|_n^2 + C\left(\sqrt{\frac{\tilde{K}_N \log^3(N)}{Nn}} + \Delta_n^2\right).$$

This is the announced result. □

Proof of Theorem 3.3.5. $i \in \mathcal{Y}$, define once again $\mathbb{E}_i = \mathbb{E}[.|\mathbb{1}_{Y_1}, \ldots, \mathbb{1}_{Y_N=i}]$. On the event $\{N_i > 1\}$, we have for all $A_i > 0$

$$\begin{aligned}
\mathbb{E}_i\left[\left\|\widehat{b}_i - b_i^*\right\|_n^2\right] &= \mathbb{E}_i\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widehat{b}_i(X_{k\Delta}) - b_i^*(X_{k\Delta})\right)^2\right] \\
&= \mathbb{E}_i\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widehat{b}_i(X_{k\Delta}) - b_i^*(X_{k\Delta})\right)^2 \mathbb{1}_{\{|X_{k\Delta}| \leq A_i\}}\right] \\
&\quad + \mathbb{E}_i\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widehat{b}_i(X_{k\Delta}) - b_i^*(X_{k\Delta})\right)^2 \mathbb{1}_{\{|X_{k\Delta}| > A_i\}}\right]. \qquad (3.49)
\end{aligned}$$

We bound each term of the *r.h.s.* of the above inequality. From Lemma 3.7.3, and Cauchy-Schwarz Inequality, under Assumption 3.2.1, we have for the second term of (3.49),

$$\mathbb{E}_i\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widehat{b}_i(X_{k\Delta}) - b_i^*(X_{k\Delta})\right)^2 \mathbb{1}_{\{|X_{k\Delta}| > A_i\}}\right] \leq C\sqrt{\exp\left(-CA_i^2\right)}. \qquad (3.50)$$

For the first term of (3.49), we observe that

$$\begin{aligned}
\mathbb{E}_i\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widehat{b}_i(X_{k\Delta}) - b_i^*(X_{k\Delta})\right)^2 \mathbb{1}_{\{|X_{k\Delta}| \leq A_i\}}|\mathcal{D}_N\right] &= \int_{-A_i}^{A_i}\left(\widehat{b}_i(x) - b_i^*(x)\right)^2\left(\frac{1}{n}\sum_{k=1}^{n-1} p(k\Delta, x)\right)\mathrm{d}x \\
&\quad + \frac{1}{n}\left(\widehat{b}_i(0) - b_i^*(0)\right)^2.
\end{aligned}$$

For $A_i = (\log(N_i))^{1/4}$ and from Lemma 3.7.4, we then deduce that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widehat{b}_i(X_{k\Delta}) - b_i^*(X_{k\Delta})\right)^2 \mathbb{1}_{N_i>1}\mathbb{1}_{\{|X_{k\Delta}| \leq A_i\}}|\mathcal{D}_N\right] \\
&\qquad \leq C_1 e^{C_2\sqrt{\log(N)}}\mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1}\left(\widehat{b}_i(X_{k\Delta}) - b_i^*(X_{k\Delta})\right)^2 \mathbb{1}_{\{|X_{k\Delta}| \leq A_i\}}|\mathcal{D}_N, Y = i\right] \\
&\qquad \leq C_1 e^{C_2\sqrt{\log(N)}}\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,i}^2\right].
\end{aligned}$$

From the above key equation, Equation (3.49), Equation (3.50), and Theorem 3.3.4, we deduce,

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_n^2\right] \leq C\left(\exp\left(C_2\sqrt{\log(N)}\right)\left(\frac{\log(N)^4}{N}\right)^{1/5} + \mathbb{E}\left[\exp(-C_2\sqrt{\log(N_i)})\mathbb{1}_{N_i>1}\right]\right).$$

Since $\exp(-C_2\sqrt{\log(N_i)})\mathbb{1}_{N_i>1} \longrightarrow 0$ *a.s.* as $N \to \infty$, and $\exp(-C_2\sqrt{\log(N_i)}) \leq 1$ for almost all $N_i > 1$, the theorem of dominated convergence implies

$$\mathbb{E}\left[\exp(-C_2\sqrt{\log(N_i)})\mathbb{1}_{N_i>1}\right] \longrightarrow 0 \text{ as } N \to \infty.$$

Besides from Theorem 3.3.4, we also have

$$\mathbb{E}\left[\left\|\widehat{\sigma}^2 - \sigma^{2*}\right\|_n\right] \leq \left(\frac{\log(N)^4}{N}\right)^{1/5}.$$

Therefore, applying Theorem 3.3.2 with $b_{\max} = \log(N)^{3/2}$, $\sigma_0^{-2} = \log(N)$, we get the desired result. □

Proof of Proposition 3.3.7. For all $i \in \mathcal{Y}$, let $\mathbb{P}_i = \mathbb{P}(.|Y = i)$ and denote by $\mathbb{P}_0$ the probability measure under which the diffusion process $X = (X_t)_{t\geq 0}$ is solution of $dX_t = d\widetilde{W}_t$ where $\widetilde{W}$ is a Brownian motion under $\mathbb{P}_0$. We deduce from the Girsanov's Theorem (see e.g. [61], Chapter III) that

$$\forall i \in \mathcal{Y},\ \forall t \in [0,1],\ \ \frac{d\mathbb{P}_i}{d\mathbb{P}_0}|_{\mathcal{F}_t^X} = \exp\left(\int_0^t b_i^*(X_s)dX_s - \frac{1}{2}\int_0^t b_i^{*2}(X_s)ds\right),$$

where $(\mathcal{F}_t^X)_{t\in[0,1]}$ is the natural filtration of X. Then, for all $i, j \in \mathcal{Y}$ such that $i \neq j$,

$$\forall t \in [0,1],\ \ \frac{d\mathbb{P}_i}{d\mathbb{P}_j}|_{\mathcal{F}_t^X} = \exp\left(\int_0^t (b_i^* - b_j^*)(X_s)dX_s - \frac{1}{2}\int_0^t (b_i^{*2} - b_j^{*2})(X_s)ds\right) \leq C\exp\left(M_t^{i,j}\right), \quad (3.51)$$

where the constant C depends on C_{b^*} given in Assumption 3.3.6 and

$$\forall i, j \in \mathcal{Y} : i \neq j,\ \ M_t^{i,j} = \int_0^t (b_i^* - b_j^*)(X_s)dW_s,\ \ t \in [0,1].$$

Then, for all $i, j \in \mathcal{Y}$ such that $i \neq j$ and for all $a > 0$, since $\left\|\widehat{b}_i - b_i^*\right\|_\infty^2 \leq 2A^2\log(N)$, and using Equation (3.51) we have

$$\begin{aligned}
\left\|\widehat{b}_i - b_i^*\right\|_{n,i}^2 &= \frac{1}{n}\sum_{k=0}^{n-1}\mathbb{E}_{X|Y=i}\left[\left(\widehat{b}_i - b_i^*\right)^2(X_{k\Delta})\right] = \frac{1}{n}\sum_{k=0}^{n-1}\mathbb{E}_{X|Y=j}\left[\left(\widehat{b}_i - b_i^*\right)^2(X_{k\Delta})\frac{d\mathbb{P}_i}{dP_j}|\mathcal{F}_{k\Delta}^X\right] \\
&\leq \frac{C}{n}\sum_{k=0}^{n-1}\mathbb{E}_{X|Y=j}\left[\left(\widehat{b}_i - b_i^*\right)^2(X_{k\Delta})\exp\left(M_{k\Delta}^{i,j}\right)\right] \\
&\leq C\exp(a)\left\|\widehat{b}_i - b_i^*\right\|_{n,j}^2 + CA^2\log(N)\mathbb{E}_{X|Y=j}\left[\exp\left(M_{k\Delta}^{i,j}\right)\mathbb{1}_{M_{k\Delta}^{i,j}>a}\right].
\end{aligned}$$

Using the Cauchy-Schwarz inequality and *Lemma 2.1 in [91]*, there exist constants $C > 0$ and $c > 0$ depending on $C_{\mathbf{b}^*}$ such that,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(M_t^{i,j}\right)\mathbb{1}_{M_t^{i,j}>a}\right] &\leq \sqrt{\mathbb{P}\left(M_t^{i,j} > a\right)}\sqrt{\mathbb{E}\left[\exp\left(2M_t^{i,j} - 2\left\langle M^{i,j}, M^{i,j}\right\rangle_t\right)\exp\left(2\left\langle M^{i,j}, M^{i,j}\right\rangle_t\right)\right]} \\
&\leq C\exp(-a^2/2c)\sqrt{\mathbb{E}\left[\exp\left(2M_t^{i,j} - 2\left\langle M^{i,j}, M^{i,j}\right\rangle_t\right)\right]},
\end{aligned}$$

where $\mathbb{P}\left(M_t^{i,j} > a\right) \leq \exp(-a^2/c)$ ([91]) and $\exp\left(2\left\langle M^{i,j}, M^{i,j}\right\rangle_t\right) < \infty\ a.s$ since the drift functions are bounded. Moreover, since $(M_t^{i,j})_{t\leq 1}$ is a martingale and

$$\mathbb{E}\left[\exp\left(\left\langle M^{i,j}, M^{i,j}\right\rangle_1\right)\right] < \infty,$$

according to the Novikov assumption, thus $\mathcal{E}(M^{i,j}) := \left\{\exp\left(2M_t^{i,j} - 2\left\langle M^{i,j}, M^{i,j}\right\rangle_t\right)\right\}_{t\leq 1}$ is a martingale with respect to the natural filtration $\mathcal{F}^M$ of $M^{i,j}$ (see [66], Proposition 5.8 and Theorem 5.9). We deduce that for all $t \in [0,1]$,

$$\mathbb{E}\left[\exp\left(2M_t^{i,j} - 2\left\langle M^{i,j}, M^{i,j}\right\rangle_t\right)\right] = \mathbb{E}\left[\mathbb{E}\left(\mathcal{E}(M^{i,j})_t|\mathcal{F}_0^M\right)\right] = \mathbb{E}\left[\exp\left(2M_0^{i,j} - 2\left\langle M^{i,j}, M^{i,j}\right\rangle_0\right)\right] = 1.$$

Thus, for all $a > 0$, we obtain $\mathbb{E}\left[\exp\left(M_t^{i,j}\right)\mathbb{1}_{M_t^{i,j}>a}\right] \leq C\exp(-a^2/c)$. Finally, set $a = \sqrt{c\log(N)}$, it follows that for all $i, j \in \mathcal{Y}$ such that $i \neq j$, there exists a constant $C > 0$ such that

$$\left\|\widehat{b}_i - b_i^*\right\|_{n,j}^2 \leq C\exp\left(\sqrt{c\log(N)}\right)\left\|\widehat{b}_i - b_i^*\right\|_{n,i}^2 + C\frac{A^2\log(N)}{N}.$$

□

Proof of Theorem 3.3.8 . From Theorem 3.3.2, and its assumptions, we have

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\left(\sqrt{\Delta} + \frac{1}{\mathfrak{p}_0^*\sqrt{N}} + \mathbb{E}\left[b_{\max}\sigma_0^{-2}\sum_{i=1}^{K}\|\widehat{b}_i - b_i^*\|_n\right] + \mathbb{E}\left[\sigma_0^{-2}\|\widehat{\sigma}^2 - \sigma^{2*}\|_n\right]\right).$$

For all $i \in \mathcal{Y}$ we obtain from Proposition 3.3.7 with $A_{N_i} = \log(N_i) \leq \log(N)$ that there exist constants $C_1, c > 0$ such that

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_n\right] = \sum_{j=1}^{K}\mathfrak{p}_j^*\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,j}\right] \leq C_1\exp\left(\sqrt{c\log(N)}\right)\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_{n,i}\right] + C_1\frac{\log^3(N)}{N}.$$

Then, from Theorem 3.3.4 with $A_{N_i} = \log(N_i)$, $K_{N_i} \propto (N_i\log(N_i))^{1/5}$ on the event $\{N_i > 1\}$ for each $i \in \mathcal{Y}$, and $\tilde{A}_N = \log(N)$, $\tilde{K}_N \propto (N\log(N))^{1/5}$ and $\Delta = \mathrm{O}(1/N)$, there exist constants $C_2, C_3 > 0$ such that

$$\forall i \in \mathcal{Y},\ \mathbb{E}\left\|\widehat{b}_i - b_i^*\right\|_{n,i} \leq C_2\left(\frac{\log^4(N)}{N}\right)^{1/5},\ \text{and}\ \mathbb{E}\left\|\widehat{\sigma}^2 - \sigma^{*2}\right\|_n \leq C_3\left(\frac{\log^4(N)}{N}\right)^{1/5}.$$

Finally, by (3.12), we deduce that there exist constants $C, c > 0$ such that

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\exp\left(\sqrt{c\log(N)}\right)N^{-1/5}.$$

□

Let us now turn to the proof of Theorem 3.4.3. We have the following lemma.

Lemma 3.7.7. *Let $\beta \geq 1$ be a real number and suppose that $K_{N_i} = \mathrm{O}\left(\log^{-5/2}(N_i)N_i^{1/(2\beta+1)}\right)$ with N_i a.s large enough, and $A_{N_i} = \sqrt{\frac{3\beta}{2\beta+1}\log(N_i)}$. Under Assumption 3.2.1, the following holds:*

$$\mathbb{P}_i\left(\Omega^c_{n,N_i,K_{N_i}}\right) \leq c\frac{K_{N_i}}{N_i},$$

where $c > 0$ is a constant.

3.7.3 Proofs of Section 3.4

Proof of Theorem 3.4.3. Note that throughout the proof we work conditional on the random variables $\mathbb{1}_{Y_1=i}, \cdots, \mathbb{1}_{Y_N=i}$ and on the event $\{N_i > 1\}$, so that N_i can be viewed as a deterministic variable. Then, to alleviate the notations, let use denote

$$\mathbb{P}_i := \mathbb{P}(.|\mathbb{1}_{Y_1=i}, \cdots, \mathbb{1}_{Y_N=i})\ \text{ and }\ \mathbb{E}_i = \mathbb{E}[.|\mathbb{1}_{Y_1=i}, \cdots, \mathbb{1}_{Y_N=i}].$$

For each class $i \in \mathcal{Y}$, the drift function b_i^* is the solution of the following regression model

$$Z_{k\Delta}^j = b_i^*(X_{k\Delta}^j) + \xi_{k\Delta}^j + R_{k\Delta}^j,\ \ j \in \mathcal{I}_i,\ \ k \in [\![0, n-1]\!],$$

where we recall that $\mathcal{I}_i$ is the set of indices j such that $Y_j = i$, and

$$\xi_{k\Delta}^j := \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}\sigma^*(X_s^j)dW_s^j,\quad R_{k\Delta}^j := \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}(b_i^*(X_s^j) - b_i^*(X_{k\Delta}^j))ds. \tag{3.52}$$

We first focus on the error $\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\right]$ for each label $i \in \mathcal{Y}$. Therefore, we consider the following decomposition:

$$\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\right] = \mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\mathbb{1}_{\Lambda_i}\right] + \mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\mathbb{1}_{\Lambda_i'}\right], \tag{3.53}$$

where

$$\Lambda_i = \Omega_{n,N_i,K_{N_i}}\ \text{ and }\ \Lambda_i' = \Omega^c_{n,N_i,K_{N_i}}.$$

Upper bound of $\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\mathbb{1}_{\Lambda_i}\right]$. From the proof of Proposition 4.4 in [30], Equation (D.5), we see that for all $h \in \mathcal{S}_{K_{N_i},M}$ and for all $a, d > 0$, we have on the event $\Lambda_i = \Omega_{n,N_i,K_{N_i}}$,

$$\left(1 - \frac{2}{a} - \frac{4}{d}\right)\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i} \leq \left(1 + \frac{2}{a} + \frac{4}{d}\right)\left\|h - b^*_{A_{N_i},i}\right\|^2_{n,N_i} + d\sup_{\left\{h\in\mathcal{S}_{K_{N_i},M},\|h\|_{n,i}=1\right\}}\nu^2(t) + aC\Delta,$$

where $C > 0$ is a constant and where for all $h \in \mathcal{S}_{K_{N_i},M}$,

$$\nu(h) = \frac{1}{N_i n}\sum_{j\in I_i}\sum_{k=0}^{n-1} h(X^j_{k\Delta})\xi^j_{k\Delta}. \tag{3.54}$$

We set $a = d = 8$, and we obtain,

$$\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\mathbb{1}_{\Lambda_i}\right] \leq 7\inf_{h\in\mathcal{S}_{K_{N_i},M}}\left\|h - b^*_{A_{N_i},i}\right\|^2_{n,i} + 32\mathbb{E}_i\left[\sup_{\left\{h\in\mathcal{S}_{K_{N_i},M},\|h\|_{n,i}=1\right\}}\nu^2(h)\right] + 32C\Delta.$$

For $h \in \mathcal{S}_{K_{N_i},M}$, $h = \sum_{\ell=-M}^{K_{N_i}-1} w_\ell B_{\ell,M,\mathbf{u}}$ and $\|h\|^2_{n,i} = w'\Psi^i_{K_{N_i}}w$ equals to one here, then $w = \Psi^{-1/2}_{K_{N_i}}u$ where the vector u satisfies $\|u\|_{2,K_{N_i}+M} = 1$. Finally, one obtains,

$$h = \sum_{\ell=-M}^{K_{N_i}-1} w_\ell B_{\ell,M,\mathbf{u}} = \sum_{\ell=-M}^{K_{N_i}-1} u_\ell\left(\sum_{\ell'=-M}^{K_{N_i}-1}\left[\Psi^{-1/2}_{K_{N_i}}\right]_{\ell',\ell} B_{\ell',M,\mathbf{u}}\right). \tag{3.55}$$

For all $h \in \mathcal{S}_{K_{N_i},M}$ such that $\|h\|_{n,i} = 1$, using Equation (3.54) and (3.55), gives

$$\nu^2(h) = \left(\sum_{\ell=-M}^{K_{N_i}-1} u_\ell\frac{1}{N_i n}\sum_{j=1}^{N_i}\sum_{k=0}^{n-1}\sum_{\ell'=-M}^{K_{N_i}-1}\left[\Psi^{-1/2}_{K_{N_i}}\right]_{\ell',\ell} B_{\ell',M,\mathbf{u}}(X^{ij}_{k\Delta})\xi^{ij}_{k\Delta}\right)^2.$$

Cauchy-Schwarz inequality together with $\|u\|_2 = 1$, produce

$$\nu^2(h) \leq \sum_{\ell=-M}^{K_{N_i}-1}\left(\frac{1}{N_i n}\sum_{j=1}^{N_i}\sum_{k=0}^{n-1}\sum_{\ell'=-M}^{K_{N_i}-1}\left[\Psi^{-1/2}_{K_{N_i}}\right]_{\ell',\ell} B_{\ell',M,\mathbf{u}}(X^{ij}_{k\Delta})\xi^{ij}_{k\Delta}\right)^2.$$

Finally we obtain,

$$\begin{aligned}\mathbb{E}_i\left[\sup_{h\in\mathcal{S}_{K_{N_i},M},\|h\|_{n,i}=1}\nu^2(h)\right] &\leq \frac{1}{N_i}\mathbb{E}_i\left[\frac{1}{n^2}\sum_{\ell=-M}^{K_{N_i}-1}\left(\sum_{k=0}^{n-1}\sum_{\ell'=-M}^{K_{N_i}-1}\left[\Psi^{-1/2}_{K_{N_i}}\right]_{\ell',\ell} B_{\ell',M,\mathbf{u}}(X^{i_1}_{k\Delta})\xi^{i_1}_{k\Delta}\right)^2\right]\\ &= \frac{1}{N_i}\mathbb{E}_i\left[\frac{1}{n^2}\sum_{\ell=-M}^{K_{N_i}-1}\sum_{k=0}^{n-1}\left(\sum_{\ell'=-M}^{K_{N_i}-1}\left[\Psi^{-1/2}_{K_{N_i}}\right]_{\ell',\ell} B_{\ell',M,\mathbf{u}}(X^{i_1}_{k\Delta})\right)^2\left(\xi^{i_1}_{k\Delta}\right)^2\right].\end{aligned}$$

According to Equation (3.52) and considering the natural filtration $(\mathcal{F}_t)_{t\geq 0}$ of the Brownian motion, for all $k \in [\![0, n-1]\!]$, we have $\mathbb{E}_i\left(\xi^{i_1}_{k\Delta}|\mathcal{F}_{k\Delta}\right) = 0$ and

$$\mathbb{E}_i\left[\left(\xi^{i_1}_{k\Delta}\right)^2|\mathcal{F}_{k\Delta}\right] = \frac{1}{\Delta^2}\mathbb{E}\left[\sigma^{*2}\left(X^{i_1}_{k\Delta}\right)\mathbb{E}\left(\left(\int_{k\Delta}^{(k+1)\Delta}\sigma^*(X^{i_1}_s)\right)^2|\mathcal{F}_{k\Delta}\right)\right] \leq \frac{\sigma^{*2}_1}{\Delta}.$$

By definition of the Gram matrix $\Psi_{K_{N_i}}$, we deduce that

$$\begin{aligned}\mathbb{E}_i\left[\sup_{h\in\mathcal{S}_{K_{N_i},M},\|h\|_{n,i}=1}\nu^2(h)\right] \leq & \frac{\sigma_1^{*2}}{N_i}\mathbb{E}_i\left[\frac{1}{n}\sum_{\ell=-M}^{K_{N_i}-1}\sum_{k=0}^{n-1}\left(\sum_{\ell'=-M}^{K_{N_i}-1}\left[\Psi_{K_{N_i}}^{-1/2}\right]_{\ell',\ell}B_{\ell',M,\mathbf{u}}(X_{k\Delta}^{1,i})\right)^2\right]\\ \leq & \frac{\sigma_1^{*2}}{N_i}\mathbb{E}_i\left(\sum_{\ell,\ell',\ell''=-M}^{K_{N_i}-1}\left[\Psi_{K_{N_i}}^{-1/2}\right]_{\ell',\ell}\left[\Psi_{K_{N_i}}^{-1/2}\right]_{\ell'',\ell}\left[\Psi_{K_{N_i}}\right]_{\ell',\ell''}\right)\\ = & \frac{\sigma_1^{*2}}{N_i}\mathbb{E}_i\left(\mathrm{Tr}\left(\Psi_{K_{N_i}}^{-1}\Psi_{K_{N_i}}\right)\right).\end{aligned}$$

Besides,

$$\mathrm{Tr}\left(\Psi_{K_{N_i}}^{-1}\Psi_{K_{N_i}}\right)=K_{N_i}+M.$$

Thus, finally, there exists a constant $C_1>0$ depending on σ_1^* and M such that

$$\mathbb{E}_i\left[\sup_{h\in\mathcal{S}_{K_{N_i},M},\|h\|_{n,i}=1}\nu^2(h)\right]\leq C_1\frac{K_{N_i}}{N_i}.$$

Thus, there exists a constant $C>0$ such that,

$$\mathbb{E}_i\left[\left\|\widehat{b}_i-b^*_{A_{N_i},i}\right\|_{n,N_i}^2\mathbb{1}_{\Lambda_i}\right]\leq 7\inf_{h\in\mathcal{S}_{K_{N_i},M}}\left\|h-b^*_{A_{N_i},i}\right\|_{n,i}^2+C\left(\frac{K_{N_i}}{N_i}+\Delta\right). \tag{3.56}$$

Upper bound of $\mathbb{E}\left[\left\|\widehat{b}_i-b^*_{A_{N_i},i}\right\|_{n,N_i}^2\mathbb{1}_{\Lambda_i'}\right]$**.** Using the Cauchy-Schwarz inequality, we have

$$\mathbb{E}_i\left[\left\|\widehat{b}_i-b^*_{A_{N_i},i}\right\|_{n,N_i}^2\mathbb{1}_{\Lambda_i'}\right]\leq C_0\log^2(N_i)\mathbb{P}_i\left(\Omega^c_{n,N_i,K_{N_i}}\right),$$

since for N large enough, using (3.8), we have,

$$\left\|\widehat{b}_i-b^*_{A_{N_i},i}\right\|_{n,N_i}^2\leq 2\|\widehat{b}_i\|_\infty^2+2\|b^*_{A_{N_i},i}\|_\infty^2\leq 4A_{N_i}^2\log(N_i)\leq C_0\log^2(N_i),$$

where $C_0>0$ is a constant. Using Lemma 2.8.9, we have

$$\mathbb{P}_i(\Lambda_i')=\mathbb{P}_i\left(\Omega^c_{n,N_i,K_{N_i}}\right)\leq c\frac{K_{N_i}}{N_i}. \tag{3.57}$$

Then, from Equation (3.57), there exists a constant $C>0$ such that

$$\mathbb{E}_i\left[\left\|\widehat{b}_i-b^*_{A_N,i}\right\|_{n,N_i}^2\mathbb{1}_{\Lambda_i'}\right]\leq C\log^2(N_i)\frac{K_{N_i}}{N_i}. \tag{3.58}$$

Upper bound of $\mathbb{E}_i\left[\left\|\widehat{b}_i-b^*_{A_{N_i},i}\right\|_{n,N_i}^2\right]$**.** From Equations (3.53), (3.56) and (3.58), there exists a constant $C>0$ such that

$$\mathbb{E}_i\left[\left\|\widehat{b}_i-b^*_{A_{N_i},i}\right\|_{n,N_i}^2\right]\leq 7\inf_{h\in\mathcal{S}_{K_{N_i},M}}\left\|h-b^*_{A_{N_i},i}\right\|_{n,i}^2+C\left(\frac{\log^2(N_i)K_{N_i}}{N_i}+\Delta\right). \tag{3.59}$$

Upper bound of $\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right]$. Using Equation (3.59), we have

$$\begin{aligned}\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right] &= \mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right] - 2\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\right] \\ &\quad + 2\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\right] \\ &\leq \mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right] - 2\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\right] \\ &\quad + 7\inf_{h\in\mathcal{S}_{K_{N_i},M}}\left\|h - b^*_{A_{N_i},i}\right\|^2_{n,i} + C\left(\frac{\log^2(N_i)K_{N_i}}{N_i} + \Delta\right).\end{aligned}$$

From the proof of Theorem 3.3.4, we deduce that

$$\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right] - 2\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,N_i}\right] \leq C\log^3(N_i)K_{N_i}/N_i$$

with $C > 0$ a constant depending on $\mathfrak{p}_0 = \min_{i\in\mathcal{Y}}\mathfrak{p}^*_i$. Besides, since $b^*_i \in \Sigma(\beta, R)$, we have

$$\inf_{h\in\mathcal{S}_{K_{N_i},M}}\|h - b^*_{A_{N_i},i}\|^2_{n,i} \leq C\left(\frac{A_{N_i}}{K_{N_i}}\right)^{2\beta},$$

where $C > 0$ is a constant (see [30], Lemma D.2). Then it comes that

$$\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right] \leq C\left(\left(\frac{A_{N_i}}{K_{N_i}}\right)^{2\beta} + \frac{K_{N_i}\log^3(N_i)}{N_i} + \Delta\right),$$

where $C > 0$ is a constant depending on β, $\Delta = \mathrm{O}(1/N)$. Since

$$K_{N_i} = \mathrm{O}\left(\log^{-5/2}(N_i)N_i^{1/(2\beta+1)}\right),$$

we obtain

$$\mathbb{E}_i\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right] \leq C\log^{6\beta}(N_i)N_i^{-\frac{2\beta}{2\beta+1}} \leq C\log^{6\beta}(N)N_i^{-\frac{2\beta}{2\beta+1}}.$$

Using the Jensen's inequality,

$$\mathbb{E}\left[\mathbb{1}_{N_i>1}\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right] \leq C\log^{6\beta}(N)\mathbb{E}\left[\mathbb{1}_{N_i>1}N_i^{-\frac{2\beta}{2\beta+1}}\right] \leq C\log^{6\beta}(N)\left(\mathbb{E}\left[\frac{\mathbb{1}_{N_i>1}}{N_i}\right]\right)^{\frac{2\beta}{2\beta+1}}.$$

Using again *Lemma 4.1* from [54], we obtain

$$\mathbb{E}\left[\mathbb{1}_{N_i>1}\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_{n,i}\right] \leq C\log^{6\beta}(N)\left(\mathbb{E}\left[\frac{\mathbb{1}_{N_i>1}}{N_i}\right]\right)^{\frac{2\beta}{2\beta+1}} \leq C\log^{6\beta}(N)N^{-\frac{2\beta}{2\beta+1}}.$$

□

Proof of Theorem 3.4.4 . For all $i \in \mathcal{Y}$, recall that $b^*_{A_{N_i},i} = b^*_i\mathbb{1}_{[-A_{N_i},A_{N_i}]}$. Furthermore, set

$$N_0 := \min_{i\in\mathcal{Y}} N_i, \quad \text{then} \quad A_{N_0} := \min_{i\in\mathcal{Y}} A_{N_i}. \tag{3.60}$$

We have

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] = \mathbb{E}\left[\left(1 - \mathcal{R}(g^*)\right)\mathbb{1}_{N_0\leq 1}\right] + \mathbb{E}\left[\left(\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right)\mathbb{1}_{N_0>1}\right].$$

Then, from Proposition 3.3.1, we deduce that

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] &\leq \sum_{i=1}^K \mathbb{P}(N_i \leq 1) + 2\sum_{i=1}^K \mathbb{E}\left[\left|\widehat{\pi}_i(X) - \pi^*_i(X)\right|\mathbb{1}_{N_0>1}\right] \\ &\leq 2KN(1-\mathfrak{p}^*_0)^{N-1} + 2\sum_{i=1}^K \mathbb{E}\left[\left|\widehat{\pi}_i(X) - \pi^*_i(X)\right|\mathbb{1}_{N_0>1}\right],\end{aligned}$$

since $\widehat{g}=1$ on the event $\{N_0 \leq 1\}$. For all $i \in \mathcal{Y}$ and on the event $\{N_0 > 1\}$,

$$|\widehat{\pi}_i(X) - \pi_i^*(X)| \leq \left|\widehat{\pi}_i(X) - \bar{\pi}_i^{A_{N_0}}(X)\right| + \left|\bar{\pi}_i^{A_{N_0}}(X) - \bar{\pi}_i^*(X)\right| + |\bar{\pi}_i^*(X) - \pi_i^*(X)|,$$

where $\bar{\pi}_i^{A_{N_0}}(X) := \phi_i\left(\bar{\mathbf{F}}^{A_{N_0}}\right)$ and $\bar{\mathbf{F}}^{A_{N_0}} = \left(\bar{F}_1^{A_{N_0}}, \cdots, \bar{F}_K^{A_{N_0}}\right)$ with

$$\forall i \in \mathcal{Y}, \quad \bar{F}_i^{A_{N_0}} = \sum_{k=0}^{n-1} b^*_{A_{N_0},i}(X_{k\Delta})(X_{(k+1)\Delta} - X_{k\Delta}) - \frac{\Delta}{2} b^{*2}_{A_{N_0},i}(X_{k\Delta}).$$

Then, there exists a constant $c > 0$ such that

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \quad \leq \quad & 2\left(\sum_{i=1}^K \mathbb{E}\left(\left|\widehat{\pi}_i(X) - \bar{\pi}_i^{A_{N_0}}(X)\right| \mathbb{1}_{N_0>1}\right) + \sum_{i=1}^K \mathbb{E}\left(\left|\bar{\pi}_i^{A_{N_0}}(X) - \bar{\pi}_i^*(X)\right| \mathbb{1}_{N_0>1}\right)\right) \\ & + c(1-\mathfrak{p}_0^*)^{N/2} + 2\sum_{i=1}^K \mathbb{E}\,|\bar{\pi}_i^*(X) - \pi_i^*(X)|.\end{aligned}$$

From the proof of Theorem 3.3.2, there exists a constant $C_1 > 0$ depending on $K, \mathfrak{p}_0^*$ and $C_{\mathfrak{b}^*}$ and a constant $C_2 > 0$ depending on K such that

$$\begin{aligned}&\sum_{i=1}^K \mathbb{E}\left|\widehat{\pi}_i(X) - \bar{\pi}_i^{A_{N_0}}(X)\right| \leq C_1\left(\frac{1}{\sqrt{N}} + \sum_{i=1}^K \mathbb{E}\left[\mathbb{1}_{N_0>1}\left\|\widehat{b}_i - b^*_{A_{N_0},i}\right\|_n\right]\right), \\ &\sum_{i=1}^K \mathbb{E}\,|\bar{\pi}_i^*(X) - \pi_i^*(X)| \leq C_2\sqrt{\Delta}.\end{aligned}$$

Thus, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \quad \leq \quad & 2C_1\left(\frac{1}{\sqrt{N}} + \sum_{i=1}^K \mathbb{E}\left[\mathbb{1}_{N_0>1}\left\|\widehat{b}_i - b^*_{A_{N_0},i}\right\|_n\right]\right) + 2C_2\sqrt{\Delta} + c(1-\mathfrak{p}_0^*)^{N/2} \\ & + 2K\sum_{i=1}^K \mathbb{E}\left[\left|\bar{F}_i^{A_{N_0}}(X) - \bar{F}_i(X)\right| \mathbb{1}_{N_0>1}\right].\end{aligned}$$

For all $i \in \mathcal{Y}$,

$$\begin{aligned}\mathbb{E}\left[\left|\bar{F}_i^{A_{N_0}}(X) - \bar{F}_i(X)\right| \mathbb{1}_{N_0>1}\right] \quad \leq \quad & \mathbb{E}\left[\left|\sum_{k=0}^{n-1} b_i^*(X_{k\Delta})\mathbb{1}_{|X_{k\Delta}|>A_{N_0}} \int_{k\Delta}^{(k+1)\Delta} b_i^*(X_s)ds\right| \mathbb{1}_{N_0>1}\right] \\ & + \frac{\Delta}{2}\sum_{k=0}^{n-1} \mathbb{E}\left[\mathbb{1}_{N_0>1} b_i^{*2}(X_{k\Delta})\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}\right] \\ & + \mathbb{E}\left|\sum_{k=0}^{n-1} b_i^*(X_{k\Delta})\mathbb{1}_{N_0>1}\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}(W_{(k+1)\Delta} - W_{k\Delta})\right|.\end{aligned}$$

Under Assumption 3.3.6, we easily obtain that

$$\mathbb{E}\left|\sum_{k=0}^{n-1} b_i^*(X_{k\Delta})\mathbb{1}_{N_0>1}\mathbb{1}_{|X_{k\Delta}|>A_{N_0}} \int_{k\Delta}^{(k+1)\Delta} b_i^*(X_s)ds\right| \leq C^2_{\mathfrak{b}^*} \sup_{t\in[0,1]} \mathbb{P}\left(\{N_0 > 1\} \cap \{|X_t| > A_{N_0}\}\right),$$

and

$$\frac{\Delta}{2}\sum_{k=0}^{n-1} \mathbb{E}\left[b_i^{*2}(X_{k\Delta})\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}\right] \leq \frac{C^2_{\mathfrak{b}^*}}{2} \sup_{t\in[0,1]} \mathbb{P}\left(\{N_0 > 1\} \cap \{|X_t| > A_{N_0}\}\right).$$

For the last term, consider the natural filtration $(\mathcal{F}_t)_{t\geq 0}$ of the Brownian motion $(W_t)_{t\geq 0}$. For all $k \in [\![0, n-1]\!]$, $X_{k\Delta}$ is measurable with respect to $\mathcal{F}_{k\Delta}$ and $W_{(k+1)\Delta} - W_{k\Delta}$ is independent of $\mathcal{F}_{k\Delta}$ since the Brownian motion is an independently increasing process. Consequently, setting,

$$\mathcal{Z} = \mathbb{E}\left|\sum_{k=0}^{n-1} b_i^*(X_{k\Delta})\mathbb{1}_{N_0>1}\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}(W_{(k+1)\Delta} - W_{k\Delta})\right|,$$

and using the Cauchy Schwarz inequality, we obtain

$$\begin{aligned}
\mathcal{Z} \leq & \left\{\mathbb{E}\left[\left(\sum_{k=0}^{n-1} b_i^*(X_{k\Delta})\mathbb{1}_{N_0>1}\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}(W_{(k+1)\Delta} - W_{k\Delta})\right)^2\right]\right\}^{1/2} \\
\leq & \left\{\mathbb{E}\left[\sum_{k,\ell=0}^{n-1} b_i^*(X_{k\Delta})b_i^*(X_{\ell\Delta})\mathbb{1}_{N_0>1}\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}\mathbb{1}_{|X_{\ell\Delta}|>A_{N_0}}(W_{(k+1)\Delta} - W_{k\Delta})(W_{(\ell+1)\Delta} - W_{\ell\Delta})\right]\right\}^{1/2} \\
\leq & \left\{2\mathbb{E}\left[\sum_{k>\ell} b_i^*(X_{k\Delta})b_i^*(X_{\ell\Delta})\mathbb{1}_{N_0>1}\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}\mathbb{1}_{|X_{\ell\Delta}|>A_{N_0}}(W_{(k+1)\Delta} - W_{k\Delta})(W_{(\ell+1)\Delta} - W_{\ell\Delta})\right]\right\}^{1/2} \\
& + \left\{\mathbb{E}\left[\sum_{k=0}^{n-1} b_i^{*2}(X_{k\Delta})\mathbb{1}_{N_0>1}\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}(W_{(k+1)\Delta} - W_{k\Delta})^2\right]\right\}^{1/2} \\
\leq & T_1 + T_2.
\end{aligned}$$

We recall that $(\mathcal{F}_t)_{t\geq 0}$ is the natural filtration of the Brownian motion $(W_t)_{t\geq 0}$. Since for all $k \in [\![0, n-1]\!]$, $X_{k\Delta}$ is $\mathcal{F}_{k\Delta}-$measurable, we have

$$\begin{aligned}
T_2^2 \leq & \mathbb{E}\left[\sum_{k=0}^{n-1} b_i^{*2}(X_{k\Delta})\mathbb{1}_{N_0>1}\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}\mathbb{E}\left[(W_{(k+1)\Delta} - W_{k\Delta})^2|\mathcal{F}_{k\Delta}\right]\right] \\
\leq & C_{\mathbf{b}^*}^2 \sup_{t\in[0,1]} \mathbb{P}(N_0 > 1, |X_t| > A_{N_0}).
\end{aligned}$$

On the other hand, for all $k, \ell \in [\![0, n-1]\!]$ such that $k > \ell$, we remark that

$$\begin{aligned}
T_1^2 \leq & 2\mathbb{E}\left[\sum_{k>\ell} b_i^*(X_{k\Delta})b_i^*(X_{\ell\Delta})\mathbb{1}_{N_0>1}\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}\mathbb{1}_{|X_{\ell\Delta}|>A_{N_0}}(W_{(\ell+1)\Delta} - W_{\ell\Delta})\mathbb{E}\left[W_{(k+1)\Delta} - W_{k\Delta}|\mathcal{F}_{k\Delta}\right]\right] \\
= & 0.
\end{aligned}$$

Thus, we deduce that

$$\mathcal{Z} = \mathbb{E}\left|\sum_{k=0}^{n-1} b_i^*(X_{k\Delta})\mathbb{1}_{N_0>1}\mathbb{1}_{|X_{k\Delta}|>A_{N_0}}(W_{(k+1)\Delta} - W_{k\Delta})\right| \leq C_{\mathbf{b}^*}\sqrt{\sup_{t\in[0,1]} \mathbb{P}(N_0 > 1, |X_t| > A_{N_0})}.$$

Finally, there exists a constant $C > 0$ such that

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\left(\frac{1}{\sqrt{N}} + \sum_{i=1}^{K}\sum_{j=1}^{K} \mathfrak{p}_j^*\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_0},i}\right\|_{n,j}\mathbb{1}_{N_0>1}\right] + \sqrt{\sup_{t\in[0,1]} \mathbb{P}(N_0 > 1, |X_t| > A_{N_0})}\right). \tag{3.61}$$

From Proposition 3.3.7 with $\alpha = 1$, for all $i, j \in \mathcal{Y}$ such that $i \neq j$, we have

$$\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_0},i}\right\|_{n,j}\mathbb{1}_{N_0>1}\right] \leq C\exp(\sqrt{c\log(N)})\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_0},i}\right\|_{n,i}\mathbb{1}_{N_0>1}\right] + C\frac{\log(N)}{N}. \tag{3.62}$$

Furthermore, for all $i \in \mathcal{Y}$, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_0},i}\right\|_{n,i} \mathbb{1}_{N_0>1}\right] &\leq \mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|_{n,i} \mathbb{1}_{N_i>1}\right] + \mathbb{E}\left[\left\|b^*_{A_{N_i},i} - b^*_{A_{N_0},i}\right\|_{n,i} \mathbb{1}_{N_0>1}\right] \\ &\leq \mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|_{n,i} \mathbb{1}_{N_i>1}\right] + \|b^*_i\|_\infty \sup_{t\in[0,1]} \mathbb{P}\left(\{A_{N_i} \geq |X_t| > A_{N_0}\} \cap \{N_0 > 1\}\right) \\ &\leq \mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|_{n,i} \mathbb{1}_{N_i>1}\right] + C_{\mathbf{b}^*} \sup_{t\in[0,1]} \sum_{j\neq i} \mathbb{P}\left(\{|X_t| > A_{N_j}\} \cap \{N_j > 1\}\right).\end{aligned}$$

We deduce from Equations (3.61) and (3.62) that there exists a constant $C > 0$ depending on $C_{\mathbf{b}^*}, K$ and p_0 such that

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C &\left(\frac{1}{\sqrt{N}} + \exp\left(\sqrt{c\log(N)}\right) \sum_{i=1}^K \mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|_{n,i} \mathbb{1}_{N_i>1}\right]\right) \\ &+ C\exp\left(\sqrt{c\log(N)}\right) \sqrt{\sup_{t\in[0,1]} \sum_{i=1}^K \mathbb{P}\left(\{|X_t| > A_{N_i}\} \cap \{N_i > 1\}\right)}.\end{aligned}$$

Under the Assumptions of the Proposition and according to Theorem 3.4.3, there exist two constants $C_1, C_2 > 0$ such that $\forall i \in \mathcal{Y}$,

$$\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|_{n,i} \mathbb{1}_{N_i>1}\right] \leq C_1 \log^{3\beta}(N) N^{-\beta/(2\beta+1)},$$

and we deduce from Lemma 3.7.5 with $q = 3/2$, for all $i \in \mathcal{Y}$, and for all $t \in [0,1]$,

$$\begin{aligned}\mathbb{P}(\{|X_t| > A_{N_i}\} \cap \{N_0 > 1\}) &= \mathbb{E}\left[\mathbb{P}(\{|X_t| > A_{N_i}\} \cap \{N_i > 1\} | \mathbb{1}_{Y_1=i}, \cdots, \mathbb{1}_{Y_N=i})\right] \\ &\leq C_2 \mathbb{E}\left[\frac{\mathbb{1}_{N_i>1}}{A_{N_i}} \exp\left(-\frac{A^2_{N_i}}{3}\right)\right].\end{aligned}$$

Thus, we obtain

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\left(\exp\left(2\sqrt{c\log(N)}\right) N^{-\beta/(2\beta+1)} + \sqrt{\sum_{i=1}^K \mathbb{E}\left[\mathbb{1}_{N_i>1} \exp\left(-\frac{A^2_{N_i}}{3}\right)\right]}\right),$$

where $C > 0$ is a constant depending on $\beta, C_{\mathbf{b}^*}, K, \mathfrak{p}^*_0$. Finally, choosing $A_{N_i} = \sqrt{\frac{6\beta}{2\beta+1}\log(N_i)}$ for each $i \in \mathcal{Y}$ leads to the attended result applying the Jensen's inequality together with Lemma 4.1 in [54]. □

Proof of Theorem 3.4.6 . From Theorem 3.3.2, as we assumed $\sigma^*(.) = 1$, the excess risk of $\widehat{g}$ satisfies

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C\left(\sqrt{\Delta} + \frac{1}{\mathfrak{p}^*_0\sqrt{N}} + \sum_{i=1}^K \mathbb{E}\left[\left\|\widehat{b}_i - b^*_i\right\|_n \mathbb{1}_{N_i>1}\right] + \sum_{i=1}^K \mathbb{P}(N_i \leq 1)\right), \tag{3.63}$$

where the constant $C > 0$ depends on $b^* = (b^*_1, \cdots, b^*_K)$ and K. For each $i \in \mathcal{Y}$, we have

$$\mathbb{P}(N_i \leq 1) \leq 2N(1 - \mathfrak{p}^*_0)^{N-1},$$

and

$$\mathbb{E}\left[\left\|\widehat{b}_i - b^*_i\right\|_n \mathbb{1}_{N_i>1}\right] \leq \sqrt{\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|^2_n \mathbb{1}_{N_i>1}\right] + \mathbb{E}\left[\left\|b^*_i \mathbb{1}_{[-A_{N_i},A_{N_i}]^c}\right\|^2_n \mathbb{1}_{N_i>1}\right]}.$$

Using the Cauchy-Schwarz inequality and Assumption 3.2.1, there exists a constant $C' > 0$ such that

$$\mathbb{E}\left[\left\|b_i^* \mathbb{1}_{[-A_{N_i},A_{N_i}]^c}\right\|_n^2 \mathbb{1}_{N_i>1}\right] \leq C' \sqrt{\sup_{t\in[0,1]} \mathbb{P}(\{|X_t| > A_{N_i}\} \cap \{N_i > 1\})}.$$

Thus, for all $i \in \mathcal{Y}$, we obtain

$$\mathbb{E}\left\|\widehat{b}_i - b_i^*\right\|_n \leq \sqrt{\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|_n^2 \mathbb{1}_{N_i>1}\right] + C' \sqrt{\sup_{t\in[0,1]} \mathbb{P}(\{|X_t| > A_{N_i}\} \cap \{N_i > 1\})}}. \tag{3.64}$$

For each label $i \in \mathcal{Y}$,

$$\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|_n^2 \mathbb{1}_{N_i>1}\right] = E\left(\mathbb{1}_{N_i>1} \int_{-A_{N_i}}^{A_{N_i}} \left(\widehat{b}_i - b^*_{A_{N_i},i}\right)^2 (x) f_{n,Y}(x) dx\right) + \frac{2\log^3(N)}{n},$$

where

$$f_{n,Y}(x) := \frac{1}{n} \sum_{k=1}^{n-1} p_{Y,X}(k\Delta, x).$$

From the proof of Lemma 3.4.2, under Assumption 3.2.1, there exist constants $C_1, C_2 > 0$ such that on the event $\{N_i > 1\}$,

$$\forall x \in [-A_{N_i}, A_{N_i}],\ f_{n,Y}(x) \geq \frac{C_1}{\log(N)} \exp\left(-\frac{2A_{N_i}^2}{3(1-\log^{-1}(N))}\right) \geq \frac{C_2}{\log(N)} \exp\left(-\frac{2}{3}A_{N_i}^2\right)\ a.s,$$

and from Lemma 3.7.4 there exists another constant $C_0 > 0$ such that $f_{n,Y}(x) \leq C_0$ for all $x \in \mathbb{R}$. Then we have

$$\forall i \in \mathcal{Y},\ \forall x \in [-A_{N_i}, A_{N_i}],\ \frac{f_{n,Y}(x)}{f_{n,i}(x)} \leq \frac{C_0}{C_2} \log(N) \exp\left(\frac{2}{3}A_N^2\right).$$

Then, for all $i \in \mathcal{Y}$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|_n^2 \mathbb{1}_{N_i>1}\right] &\leq \mathbb{E}\left[\mathbb{1}_{N_i>1} \int_{-A_{N_i}}^{A_{N_i}} \left(\widehat{b}_i - b^*_{A_{N_i},i}\right)^2 (x) f_{n,i}(x) \frac{f_{n,Y}(x)}{f_{n,i}(x)}\right] + \frac{2\log^3(N)}{n} \\
&\leq \frac{C_0}{C_2} \log(N) \exp\left(\frac{2}{3}A_N^2\right) \mathbb{E}\left[\left\|\widehat{b}_i - b^*_{A_{N_i},i}\right\|_{n,i}^2 \mathbb{1}_{N_i>1}\right] + \frac{2\log^3(N)}{n}.
\end{aligned}$$

From Theorem 3.4.3, Equation (3.64) and for $n \propto N$, there exists a constant $C_3 > 0$ such that

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_n \mathbb{1}_{N_i>1}\right] \leq C_3 \sqrt{\exp\left(\frac{2}{3}A_N^2\right) \log^{6\beta+1}(N) N^{-\frac{2\beta}{2\beta+1}} + \sqrt{\sup_{t\in[0,1]} \mathbb{P}(\{|X_t| > A_{N_i}\} \cap \{N_i > 1\})}}.$$

Using the Markov inequality, for all $t \in [0,1]$, we have

$$\begin{aligned}
\mathbb{P}(\{|X_t| > A_{N_i}\} \cap \{N_i > 1\}) &= \mathbb{E}\left[\mathbb{P}\left(\{\exp(4|X_t|^2) > \exp(4A_{N_i}^2)\} \cap \{N_i > 1\} | \mathbb{1}_{Y_1=i}, \cdots, \mathbb{1}_{Y_N=i}\right)\right] \\
&\leq \mathbb{E}\left[\exp(4|X_t|^2)\right] \mathbb{E}\left[\exp(-4A_{N_i}^2) \mathbb{1}_{N_i>1}\right],
\end{aligned}$$

and since $\sigma^*(.) = 1$ and under Assumption 3.4.5, there exists a strictly positive constant C_* such that $\mathbb{E}\left[\exp(4|X_t|^2)\right] \leq C_*$ (according to [52], Proposition 1.1). Thus, there exists a constant $C > 0$ such that

$$\mathbb{E}\left[\left\|\widehat{b}_i - b_i^*\right\|_n \mathbb{1}_{N_i>1}\right] \leq C\left(\exp\left(\frac{1}{3}A_N^2\right) \log^{3\beta+1}(N) N^{-\beta/(2\beta+1)}\right) + C\mathbb{E}\left[\exp(-4A_{N_i}^2) \mathbb{1}_{N_i>1}\right]. \tag{3.65}$$

From Equations (3.65) and (3.63), we finally obtain

$$\mathbb{E}\left[\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)\right] \leq C \log^{3\beta+1}(N) N^{-3\beta/4(2\beta+1)},$$

with $A_{N_i} \leq \sqrt{\frac{3\beta}{4(2\beta+1)} \log(N_i)}$ and $C > 1$ a new constant. □

3.8 Appendix

Proof of Lemma 3.7.1. Let $s,t \in [0,1]$ with $s<t$, and $q \geq 1$. By convexity of $x \mapsto |x|^{2q}$, we have

$$|X_t - X_s|^{2q} \leq 2^{2q-1}\left(\left|\int_s^t b_Y^*(X_u)du\right|^{2q} + \left|\int_s^t \sigma(X_u)dW_u\right|^{2q}\right).$$

Then, from Jensen's inequality, we have

$$\left|\int_s^t b_Y^*(X_u)du\right|^{2q} \leq (t-s)^{2q-1}\int_s^t |b_Y^*(X_u)|^{2q}\,du,$$

Hence, under Assumption 3.2.1 on function b_Y^*, we deduce that

$$\mathbb{E}\left[\left|\int_s^t b_Y^*(X_u)du\right|^{2q}\right] \leq C_q(t-s)^{2q}\left(1+\mathbb{E}\left[\sup_{t\in[0,1]}|X_s|^{2q}\right]\right),$$

and using Burkholder-Davis-Gundy inequality, we obtain

$$\forall m>0,\quad \mathbb{E}\left[\left(\int_s^t \sigma(X_u)dW_u\right)^{2m}\right] \leq C_m\mathbb{E}\left[\left(\int_s^t \sigma^2(X_u)du\right)^{m}\right] \leq C_m\sigma_1^{2m}(t-s)^m.$$

From the above equalities, we get

Finally, as the process has finite moments, we obtain that

$$\mathbb{E}\,|X_t - X_s|^{2q} \leq C(t-s)^q,$$

where C is a constant depending on q, L_0, and σ_1. □

Proof of Lemma 3.4.2 . For all $i \in \mathcal{Y}$ and on the event $\{N_i > 1\}$, let us consider a vector $\left(x_{-M},\cdots,x_{K_{N_i}-1}\right) \in \mathbb{R}^{K_{N_i}+M}$ such that $x_j \in [u_{j+M},u_{j+M+1})$ and $B_{j,M,\mathbf{u}}(x_j) \neq 0$. Since we have $[u_{j+M},u_{j+M+1}) \cap [u_{j'+M},u_{j'+M+1}) = \emptyset$ for all $j,j' \in \{-M,\cdots,K_{N_i}-1\}$ such that $j \neq j'$, then for all $j,j' \in \{-M,\cdots,K_{N_i}-1\}$ such that $j \neq j', B_{j,M,(u)}(x_{j'}) = 0$. Consequently, we obtain:

$$\begin{aligned}\det\left(\left(B_{\ell,M,\mathbf{u}}(x_{\ell'})\right)_{-M\leq\ell,\ell'\leq K_{N_i}-1}\right) &= \det\left(\operatorname{diag}\left(B_{-M,M,\mathbf{u}}(x_M),\cdots,B_{K_{N_i}-1,M,\mathbf{u}}(x_{K_{N_i}-1})\right)\right)\\ &= \prod_{\ell=-M}^{K_{N_i}-1} B_{\ell,M,\mathbf{u}}(x_\ell) \neq 0.\end{aligned}$$

Then, we deduce from [18], *Lemma 1* that the matrix $\Psi_{K_{N_i}}$ is invertible for all $K_{N_i} \in \mathcal{K}_{N_i}$, where the interval $[-A_{N_i},A_{N_i}]$ and the function f_T is replaced by $f_n : x \mapsto \frac{1}{n}\sum_{k=0}^{n-1} p(k\Delta,x)$ with $\lambda([-A_{N_i},A_{N_i}] \cap \operatorname{supp}(f_n)) > 0$, λ being the Lebesgue measure.

For all $w \in \mathbb{R}^{K_{N_i}+M}$ such that $\|w\|_{2,K_{N_i}+M} = 1$, we have:

$$w'\Psi_{K_{N_i}}w = \|h_w\|_n^2 = \int_{-A_{N_i}}^{A_{N_i}} h_w^2(x)f_n(x)dx + \frac{h_w^2(x_0)}{n} \quad \text{with } h_w = \sum_{\ell=-M}^{K_{N_i}-1} w_\ell B_{\ell,M,\mathbf{u}}.$$

Since $\sigma^* = 1$, according to Lemma 3.7.5, under Assumption 3.2.1, the transition density satisfies:

$$\forall (t,x) \in (0,1]\times\mathbb{R},\ \frac{1}{K_q\sqrt{t}}\exp\left(-\frac{(2q-1)x^2}{2qt}\right) \leq p_X(t,x) \leq \frac{K_q}{\sqrt{t}}\exp\left(-\frac{x^2}{2qt}\right) \text{ where } K_q>1 \text{ and } q>1.$$

We set $q = 3/2$, thus, since $s \mapsto \exp\left(-(2q-1)x^2/2qs\right)$ is an increasing function, we have on the event $\{N_i > 1\}$ and for all $x \in [-A_{N_i}, A_{N_i}]$,

$$\begin{aligned} f_n(x) &\geq \frac{1}{Cn}\sum_{k=1}^{n-1}\exp\left(-\frac{2x^2}{3k\Delta}\right) \geq \frac{1}{C}\int_0^{(n-1)\Delta}\exp\left(-\frac{2x^2}{3s}\right)ds \\ &\geq \frac{1}{C}\int_{1-\log^{-1}(N_i)}^{1-2^{-1}\log^{-1}(N_i)}\exp\left(-\frac{2x^2}{3s}\right)ds \\ &\geq \frac{1}{2C\log(N_i)}\exp\left(-\frac{2A_{N_i}^2}{3(1-\log^{-1}(N_i))}\right). \end{aligned}$$

Finally, since there exists a constant $C_1 > 0$ such that $\|h_w\|^2 \geq C_1 A_{N_i} K_{N_i}^{-1}$ (see [30], Lemma 2.6), for all $w \in \mathbb{R}^{K_{N_i}+M}$ such that $\|w\|_{2,K_{N_i}+M} = 1$, there exists constants $C', C > 0$ such that,

$$w'\Psi_{K_{N_i}}w \geq \frac{C'A_{N_i}}{K_{N_i}\log(N_i)}\exp\left(-\frac{2A_{N_i}^2}{3(1-\log^{-1}(N_i))}\right) \geq \frac{CA_{N_i}}{K_{N_i}\log(N_i)}\exp\left(-\frac{2}{3}A_{N_i}^2\right).$$

Furthermore, we set $w_0 = e_{K_{N_i}-1} \in \mathbb{R}^{K_{N_i}+M}$ where for all $\ell \in [\![-M, K_{N_i}-1]\!]$,

$$\left[e_{K_{N_i}-1}\right]_\ell := \delta_{\ell,K_{N_i}-1} = \begin{cases} 0 \text{ if } \ell \neq K_{N_i}-1 \\ 1 \text{ else.} \end{cases}$$

We have,

$$\begin{aligned} w_0'\Psi_{K_{N_i}}w_0 &= \int_{-A_{N_i}}^{A_{N_i}} B_{K_{N_i}-1,M,\mathbf{u}}^2(x)f_n(x) + \frac{B_{K_{N_i}-1,M,\mathbf{u}}(0)}{n} \\ &\leq \frac{C}{n}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k\Delta}}\exp\left(-\frac{u_{K_{N_i}-1}^2}{3k\Delta}\right)\left\|B_{K_{N_i}-1,M,\mathbf{u}}\right\|^2 + \frac{1}{n} \\ &\leq \frac{CC_1A_{N_i}K_{N_i}^{-1}}{n}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k\Delta}}\exp\left(-\frac{\alpha_{N_i}^2}{3k\Delta}\right) + \frac{1}{n}, \end{aligned}$$

where $\alpha_{N_i} = A_{N_i}(K_{N_i}-2)/K_{N_i}$, $\left\|B_{K_{N_i}-1,M,\mathbf{u}}\right\|^2 \leq C_1 A_{N_i}K_{N_i}^{-1}$ (see [30], Lemma 2.6) and $C_1 > 0$ is a constant. Since the function $s \mapsto \exp\left(-\alpha_{N_i}^2/3s\right)/\sqrt{s}$ is increasing, we deduce that

$$n^{-1}\sum_{k=1}^{n-1}\frac{1}{\sqrt{k\Delta}}\exp\left(-\alpha_{N_i}^2/3k\Delta\right) \leq n^{-1}\sum_{k=1}^{n-1}\exp\left(-\alpha_{N_i}^2/3\right),$$

and for N large enough,

$$w_0'\Psi_{K_{N_i}}w_0 \leq \frac{CA_{N_i}}{K_{N_i}}\exp\left(-\frac{A_{N_i}^2}{3}\left(\frac{K_{N_i}-2}{K_{N_i}}\right)^2\right) + \frac{1}{n} \leq \frac{C'A_{N_i}}{K_{N_i}}\exp\left(-\frac{A_{N_i}^2}{3}\left(\frac{K_{N_i}-2}{K_{N_i}}\right)^2\right),$$

where $C' > 0$ is a constant and $n \geq N \geq N_i$. □

Proof of Lemma 2.8.9 . Let us remind the reader of the Gram matrix $\Psi_{K_{N_i}}$ given in Equation (3.13) for $i \in \mathcal{Y}$,

$$\Psi_{K_{N_i}} = \mathbb{E}\left[\frac{1}{N_i n}\mathbf{B}_{K_{N_i}}'\mathbf{B}_{K_{N_i}}\right] = \mathbb{E}\left(\widehat{\Psi}_{K_{N_i}}\right),$$

where, on the event $\{N_i > 1\}$, and denoting by $\mathcal{I}_i := \{i_1, \dots, i_{N_i}\}$ the indices j such that $Y_j = i$,

$$\mathbf{B}_{K_{N_i}} := \begin{pmatrix} B_{-M}\left(X_0^{i_1}\right) & \dots & \dots & B_{K_{N_i}-1}\left(X_0^{i_1}\right) \\ \vdots & & & \vdots \\ B_{-M}\left(X_{(n-1)\Delta}^{i_1}\right) & \dots & \dots & B_{K_{N_i}-1}\left(X_{(n-1)\Delta}^{i_1}\right) \\ \vdots & & & \vdots \\ B_{-M}\left(X_0^{i_{N_i}}\right) & \dots & \dots & B_{K_{N_i}-1}\left(X_0^{i_{N_i}}\right) \\ \vdots & & & \vdots \\ B_{-M}\left(X_{(n-1)\Delta}^{i_{N_i}}\right) & \dots & \dots & B_{K_{N_i}-1}\left(X_{(n-1)\Delta}^{i_{N_i}}\right) \end{pmatrix} \in \mathbb{R}^{N_i n \times (K_{N_i}+M)}. \tag{3.66}$$

The empirical counterpart $\widehat{\Psi}$ is the random matrix given by $\widehat{\Psi}_{K_{N_i}}$ of size $(K_{N_i}+M)\times(K_{N_i}+M)$ is given by

$$\widehat{\Psi}_{K_{N_i}} := \frac{1}{N_i n}\mathbf{B}'_{K_{N_i}}\mathbf{B}_{K_{N_i}} = \left(\frac{1}{N_i n}\sum_{j=1}^{N_i}\sum_{k=0}^{n-1} B_\ell(X_{k\Delta}^{i_j})B_{\ell'}(X_{k\Delta}^{i_j})\right)_{\ell,\ell'\in[-M,K_{N_i}-1]}. \tag{3.67}$$

We build an orthonormal basis $\theta = (\theta_{-M}, \cdots, \theta_{K_{N_i}-1})$ of the subspace $\mathcal{S}_{K_{N_i},M}$ with respect to the $\mathbb{L}^2$ inner product $\langle ., . \rangle$ through the Gram-Schmidt orthogonalization of the spline basis $(B_{-M}, \cdots, B_{K_{N_i}-1})$. Then, we have

$$\mathrm{Span}(B_{-M}, \cdots, B_{K_{N_i}-1}) = \mathrm{Span}(\theta_{-M}, \cdots, \theta_{K_{N_i}-1}) = \mathcal{S}_{K_{N_i},M},$$

and the matrix given in Equation (3.66) is factorized as follows

$$\mathbf{B}_{K_{N_i}} = \mathbf{\Theta}_{K_{N_i}}\mathbf{R}_{K_{N_i}}, \tag{3.68}$$

where

$$\mathbf{\Theta}_{K_{N_i}} = \left(\left(\theta_\ell(X_0^{i_j}), \theta_\ell(X_\Delta^{i_j}), \cdots, \theta_\ell(X_{n\Delta}^{i_j})\right)'\right)_{\substack{1\le j\le N_i \\ -M\le \ell\le K_{N_i}-1}} \in \mathbb{R}^{N_i n\times(K_{N_i}+M)},$$

and $\mathbf{R}_{K_{N_i}}$ is an upper triangular matrix of size $(K_{N_i}+M)\times(K_{N_i}+M)$ see [67]). Let $\Phi_{K_{N_i}}$ be the Gram matrix under the orthonormal basis $\theta = \left(\theta_{-M}, \cdots, \theta_{K_{N_i}-1}\right)$ and given by

$$\Phi_{K_{N_i}} = \mathbb{E}\left[\frac{1}{N_i n}\mathbf{\Theta}'_{K_{N_i}}\mathbf{\Theta}_{\mathbf{K_{N_i}}}\right] = \mathbb{E}\left(\widehat{\Phi}_{K_{N_i}}\right),$$

where,

$$\widehat{\Phi}_{K_{N_i}} := \frac{1}{N_i n}\mathbf{\Theta}'_{K_{N_i}}\mathbf{\Theta}_{K_{N_i}} = \left(\frac{1}{N_i n}\sum_{j=1}^{N_i}\sum_{k=0}^{n-1}\theta_\ell(X_{k\Delta}^{i_j})\theta_{\ell'}(X_{k\Delta}^{i_j})\right)_{\ell,\ell'\in[-M,K_{N_i}-1]}. \tag{3.69}$$

The matrices $\Psi_{K_{N_i}}$ and $\widehat{\Psi}_{K_{N_i}}$ are respectively linked to the matrices $\Phi_{K_{N_i}}$ and $\widehat{\Phi}_{K_{N_i}}$ through the following relations

$$\Psi_{K_{N_i}} = \mathbf{R}'_{K_{N_i}}\Phi_{K_{N_i}}\mathbf{R}_{K_{N_i}} \text{ and } \widehat{\Psi}_{K_{N_i}} = \mathbf{R}'_{K_{N_i}}\widehat{\Phi}_{K_{N_i}}\mathbf{R}_{K_{N_i}}.$$

Since for all $h = \sum_{\ell=-M}^{K_{N_i}-1} a_\ell B_{\ell,M,\mathbf{u}} \in S_{K_{N_i},M}$ one has

$$\|h\|^2_{n,N_i} = a'\widehat{\Psi}_{K_{N_i}}a \text{ and } \|h\|^2_{n,i} = a'\Psi_{K_{N_i}}a, \text{ with } a = \left(a_{-M}, \cdots, a_{K_{N_i}-1}\right)',$$

we deduce that

$$\|h\|^2_{n,N_i} = w'\widehat{\Phi}_{K_{N_i}}w \text{ and } \|h\|^2_{n,i} = w'\Phi_{K_{N_i}}w, \text{ with } w = \mathbf{R}_{K_{N_i}}a.$$

Under Assumption 3.2.1, we follow the lines of [19] *Proposition 2.3* and *Lemma 6.2*. Then,

$$\begin{aligned}\sup_{h\in S_{K_{N_i},M},\|h\|_{n,i}=1}\left|\|h\|_{n,N_i}^2-\|h\|_{n,i}^2\right| &= \sup_{w\in\mathbb{R}^{K_{N_i}+M},\left\|\Phi_{K_{N_i}}^{1/2}w\right\|_{2,K_{N_i}+M}=1}\left|w'\left(\widehat{\Phi}_{K_{N_i}}-\Phi_{K_{N_i}}\right)w\right| \\ &= \sup_{u\in\mathbb{R}^{K_{N_i}+M},\|u\|_{2,K_{N_i}+M}=1}\left|u'\Phi_{K_{N_i}}^{-1/2}\left(\widehat{\Phi}_{K_{N_i}}-\Phi_{K_{N_i}}\right)\Phi_{K_{N_i}}^{-1/2}u\right| \\ &= \left\|\Phi_{K_{N_i}}^{-1/2}\widehat{\Phi}_{K_{N_i}}\Phi_{K_{N_i}}^{-1/2}-\mathrm{Id}_{K_{N_i}+M}\right\|_{\mathrm{op}}.\end{aligned}$$

Therefore,

$$\Omega_{n,N_i,K_{N_i}}^c=\left\{\left\|\Phi_{K_{N_i}}^{-1/2}\widehat{\Phi}_{K_{N_i}}\Phi_{K_{N_i}}^{-1/2}-\mathrm{Id}_{K_{N_i}+M}\right\|_{\mathrm{op}}>1/2\right\}.$$

Then, we apply here Theorem 1 of [16], it yields

$$\mathbb{P}_i\left(\Omega_{n,N_i,K_{N_i}}^c\right)\leq 2(K_{N_i}+M)\exp\left(-c_{1/2}\frac{N_i}{\mathcal{L}(K_{N_i}+M)(\|\Phi_{K_{N_i}}^{-1}\|_{\mathrm{op}}\vee 1)}\right), \tag{3.70}$$

with $c_{1/2}=(3\log(3/2)-1)/2$ and $\mathcal{L}(K_{N_i}+M):=\sup_{x\in[-A_{N_i},A_{N_i}]}\sum_{\ell=-M}^{K_{N_i}-1}\theta_\ell^2(x)$ (from application of Lemma 6.2 from [19]). For all $h=\sum_{\ell=-M}^{K_{N_i}-1}w_\ell\theta_\ell\in\mathrm{Span}\left(\theta_{-M},\cdots,\theta_{K_{N_i}-1}\right)=\mathcal{S}_{K_{N_i},M}$, we have

$$\|h\|^2=\|w\|_{2,K_{N_i}+M}^2 \text{ and } \|h\|_{n,i}^2=1 \text{ implies } w=\Phi_{K_{N_i}}^{-1/2}u \text{ where } u\in\mathbb{R}^{K_{N_i}+M}:\|u\|_{2,K_{N_i}+M}=1.$$

We deduce that

$$\sup_{h\in\mathcal{S}_{K_{N_i}+M},\,\|h\|_{n,i}^2=1}\|h\|^2=\sup_{u\in\mathbb{R}^{K_{N_i}+M},\|u\|_{2,K_{N_i}+M}=1}u'\Phi_{K_{N_i}}^{-1}u=\left\|\Phi_{K_{N_i}}^{-1}\right\|_{\mathrm{op}}.$$

Furthermore, for all $h=\sum_{\ell=-M}^{K_{N_i}-1}a_\ell B_\ell\in\mathrm{Span}\left(B_{-M},\cdots,B_{K_{N_i}-1}\right)=\mathcal{S}_{K_{N_i},M}$, we have on one side

$$\|h\|_{n,i}^2=1 \text{ implies } a=\Psi_{K_{N_i}}^{-1/2}u \text{ where } u\in\mathbb{R}^{K_{N_i}+M}:\|u\|_{2,K_{N_i}+M}=1,$$

and on the other side, for all $h\in\mathcal{S}_{K_{N_i}+M}$ such that $\|h\|_{n,i}^2=1$, from [30] *Lemma 2.6*, there exists a constant $C>0$ such that,

$$\|h\|^2\leq CA_{N_i}K_{N_i}^{-1}\|a\|_{2,K_{N_i}+M}^2=CA_{N_i}K_{N_i}^{-1}u'\Psi_{K_{N_i}}^{-1}u.$$

Then we have $a.s$

$$\left\|\Phi_{K_{N_i}}^{-1}\right\|_{\mathrm{op}}=\sup_{h\in\mathcal{S}_{K_{N_i}+M},\,\|h\|_{n,i}^2=1}\|h\|^2\leq\frac{CA_{N_i}}{K_{N_i}}\sup_{u\in\mathbb{R}^{K_{N_i}+M},\|u\|_{2,K_{N_i}+M}=1}u'\Psi_{K_{N_i}}^{-1}u=\frac{CA_{N_i}}{K_{N_i}}\left\|\Psi_{K_{N_i}}^{-1}\right\|_{\mathrm{op}}. \tag{3.71}$$

From Equations (3.70) and (3.71), there exists a constant $C>0$ such that

$$\mathbb{P}_i\left(\Omega_{n,N_i,K_{N_i}}^c\right)\leq 2(K_{N_i}+M)\exp\left(-C\frac{N_iK_{N_i}}{A_{N_i}\mathcal{L}(K_{N_i}+M)\left\|\Psi_{K_{N_i}}^{-1}\right\|_{\mathrm{op}}}\right). \tag{3.72}$$

We have $\mathcal{L}(K_{N_i}+M):=\sup_{x\in[-A_{N_i},A_{N_i}]}\sum_{\ell=-M}^{K_{N_i}-1}\theta_\ell^2(x)$ and the functions $\theta_\ell,\ \ell=-M,\ldots,K_{N_i}-1$ are given by

$$\theta_{-M}=\frac{f_{-M}}{\|f_{-M}\|} \text{ with } f_{-M}=B_{-M}$$

$$\theta_\ell=\frac{f_\ell}{\|f_\ell\|} \text{ with } f_\ell=B_\ell-\sum_{k=-M}^{\ell-1}\langle B_\ell,\theta_k\rangle\,\theta_k,\ \ \ell=-M+1,\ldots,K_{N_i}-1.$$

Note that for all $x \in [-A_{N_i}, A_{N_i}]$, there exists $\ell \in [\![-M, K_{N_i}-1]\!]$ such that $x \in [u_\ell, u_{\ell+1})$. Then, $x \in [u_{\ell'}, u_{\ell'+M+1})$ for all $\ell' \in [\![\ell - M, \ell]\!]$ if $\ell \geq 0$ and $\ell' \in [\![-M, \ell]\!]$ for $\ell \leq -1$. Thus, for each $x \in [-A_{N_i}, A_{N_i}]$, there exists at most $M+1$ spline functions that don't vanish at x. As a result, we have on one side,

$$\forall\, x \in [-A_{N_i}, A_{N_i}], \quad \sum_{\ell=-M}^{K_{N_i}-1} \theta_\ell^2(x) = \sum_{j=1}^{M+1} \theta_{\ell_j}^2(x), \tag{3.73}$$

where for all $x \in [-A_{N_i}, A_{N_i}]$, there exists integers $\ell_j,\ j \in [\![1, M+1]\!]$ such that

$$x \in \bigcap_{j=1}^{M+1} [u_{\ell_j}, u_{\ell_j+M+1}) \ \text{ and } \ x \notin [-A_{N_i}, A_{N_i}] \setminus \bigcap_{j=1}^{M+1} [u_{\ell_j}, u_{\ell_j+M+1}).$$

One the other side, for all $\ell \in [\![1, K_{N_i}-1]\!]$ and for all $x \in [u_\ell, u_{\ell+M+1})$ there exists at most $M+1$ integers $\ell_1, \ldots, \ell_{M+1}$ such that

$$\theta_\ell(x) = \frac{f_\ell(x)}{\|f_\ell\|} \ \text{ and } \ f_\ell(x) = B_\ell(x) - \sum_{j=1}^{M+1} \langle B_\ell, \theta_{\ell_j} \rangle\, \theta_{\ell_j}(x).$$

Now we focus on the supremum norm of each basis function $\theta_\ell,\ \ell = -M, \ldots, K_{N_i}-1$. For all each $\ell \in [\![-M, K_{N_i}-1]\!]$, since the spline function B_ℓ is non-zero, positive and continuous on the interval $[u_\ell, u_{\ell+M+1})$, there exists an interval $[\alpha_\ell, \beta_\ell] \subset [u_\ell, u_{\ell+M+1})$ such that $c_\ell = \inf_{x \in [\alpha_\ell, \beta_\ell]} B_\ell(x) > 0$ where $(\alpha_\ell - \beta_\ell) \propto A_{N_i}/K_{N_i}$ since $\int_{\alpha_\ell}^{\beta_\ell} B_\ell(x)dx \propto A_{N_i}/K_{N_i}$. Then we have

$$\forall\, \ell \in [\![-M, K_{N_i}-1]\!], \ \ \|B_\ell\|^2 = \int_{u_\ell}^{u_{\ell+M+1}} B_\ell^2(x)dx \geq c_\ell \int_{\alpha_\ell}^{\beta_\ell} B_\ell(x)dx = C_\ell \frac{A_{N_i}}{K_{N_i}}, \tag{3.74}$$

where the constant $C_\ell > 0$ depends on $c_\ell = \inf_{x \in [\alpha_\ell, \beta_\ell]} B_\ell(x) > 0$. Then, for $\ell = -M$, there exists a constant C_{-M} such that $\theta_{-M}^2(x) \leq C_{-M} K_{N_i}$ and for each $\ell \geq -M+1$, since the function f_ℓ depends on splines functions $B_{-M}, \ldots, B_\ell$ and only B_ℓ does not vanish on the interval $[u_{\ell+M}, u_{\ell+M+1})$, we obtain that

$$\|f_\ell\|^2 = \int_{-A_{N_i}}^{A_{N_i}} f_\ell^2(x) \geq \int_{u_{\ell+M}}^{u_{\ell+m+1}} B_\ell^2(x)dx.$$

Moreover, since B_ℓ is non-zero, positive and continue on the interval $[u_{\ell+M}, u_{\ell+M+1})$, there exists an interval $[\alpha_\ell, \beta_\ell] \subset [u_{\ell+M}, u_{\ell+M+1})$ with $(\alpha_\ell - \beta_\ell) \propto A_{N_i}/K_{N_i}$ such that $c_\ell = \inf_{x \in [\alpha_\ell, \beta_\ell]} B_\ell(x) > 0$. Then we obtain

$$\|f_\ell\|^2 \geq c_\ell \int_{\alpha_\ell}^{\beta_\ell} B_\ell(x)dx = C \frac{A_{N_i}}{K_{N_i}}, \ \ \ell \in [\![-M+1, \ldots, K_{N_i}-1]\!], \tag{3.75}$$

where $C > 0$ is a constant depending on $\min_{\ell=-M+1,\ldots,K_{N_i}-1} c_\ell > 0$. On the other side, for all $\ell \in [\![-M+1, K_{N_i}-1]\!]$ and for all $x \in [-A_{N_i}, A_{N_i}]$,

$$|f_\ell(x)| \leq |B_\ell(x)| + \sum_{j=-M}^{\ell-1} \frac{\langle B_\ell, f_j \rangle}{\|f_j\|^2} |f_j(x)| \leq 1 + C \sum_{j=-M}^{\ell-1} |f_j(x)|, \tag{3.76}$$

where the constant $C > 0$ is the upper-bound of $\langle B_\ell, f_j \rangle / \|f_j\|^2 \leq \|B_\ell\|/\|f_j\|$ according to Equations (3.74) and (3.75). For $\ell = -M$, we have $\|f_{-M}\|_\infty < \infty$. Let $\ell \in [\![-M+1, K_{N_i}-1]\!]$. Assume that the functions $f_{-M}, \ldots, f_{\ell-1}$ are all bounded, then by recurrence hypothesis, we have from Equation (3.76) that

$$\|f_\ell\|_\infty \leq 1 + C \sum_{j=-M}^{\ell-1} \|f_j\|_\infty < \infty.$$

Thus, we obtain by recurrence that the functions $f_\ell,\ \ell = -M, \dots, K_{N_i} - 1$ are bounded and finally conclude from Equation (3.73) that

$$\mathcal{L}(K_{N_i} + M) := \sup_{x \in [-A_{N_i}, A_{N_i}]} \sum_{\ell=-M}^{K_{N_i}-1} \theta_\ell^2(x) \leq C K_{N_i},$$

where the constant $C > 0$ depends on the spline basis. We deduce from Equation (3.72) that there exists a constant $C > 0$ such that

$$\mathbb{P}_i \left(\Omega^c_{n,N_i,K_{N_i}} \right) \leq 2(K_{N_i} + M) \exp \left(-C \frac{N_i}{A_{N_i} \left\| \Psi^{-1}_{K_{N_i}} \right\|_{\text{op}}} \right). \tag{3.77}$$

Furthermore, since $A_{N_i} \leq \sqrt{\frac{3\beta}{2\beta+1} \log(N_i)}$, $K_{N_i} \propto \log^{-5/2}(N_i) N_i^{1/(2\beta+1)}$ and from Lemma 3.4.2, we obtain from Equation (3.77),

$$\mathbb{P}_i \left(\Omega^c_{n,N_i,K_{N_i}} \right) \leq 2(K_{N_i} + M) \exp \left(-C \log^{3/2}(N_i) \right), \tag{3.78}$$

where $C > 0$ is a new constant depending on C_θ, β and M. Since $N_i \longrightarrow \infty\ a.s.$ as $N \longrightarrow \infty$, one has

$$\exp \left(\log(N_i) - C \log^{3/2}(N_i) \right) \longrightarrow 0\ a.s.\ \text{ as }\ N \longrightarrow \infty.$$

Then, for N large enough, $\exp \left(\log(N_i) - C \log^{3/2}(N_i) \right) \leq 1\ \ a.s.$ and from Equation (3.78),

$$\mathbb{P}_i \left(\Omega^c_{n,N_i,K_{N_i}} \right) \leq \frac{2(K_{N_i} + M)}{N_i} \leq c \frac{K_{N_i}}{N_i},$$

where the constant $c > 0$ depends on M. □

Chapter 4

Classifieur non-paramétrique de type ERM pour les trajectoires de diffusion

Résumé : Dans le contexte de la classification multiclasses de trajectoires de diffusion, nous considérons un mélange de processus de diffusion, solutions d'équations différentielles stochastiques avec un même coefficient de diffusion, et des coefficients de dérive qui dépendent des classes modélisées par une variable aléatoire discrète. Nous supposons que les coefficients des processus de diffusion et la loi discrète génèrant les étiquettes sont inconnus, et nous construisons une procédure de classification non-paramétrique basée sur la minimisation du risque empirique de classification. Nous prouvons la consistance du classifieur empirique sous des hypothèses suffisamment faibles, et nous établissons ensuite des vitesses de convergence sous un ensemble d'hypothèses sur le modèle étudié. Enfin, nous complétons nos résultats théoriques par une étude numérique à partir de données simulées.

Mots clés. Classification multiclasses, Minimisation du risque empirique, Processus de diffusion, Hypothèse de marge, Observations répétées.

Abstract : In the context of multiclass classification of diffusion paths, we consider a mixture of diffusion processes, solutions of stochastic differential equations with a common diffusion coefficient, and the drift coefficients which depend on the classes modeled by a discrete random variable. We assume that the coefficients of the diffusion processes together with the discrete law that generate the classes are unknown, and we construct a nonparametric classification procedure based on the minimization of the empirical classification risk. We prove the consistency of the empirical classifier under mild assumptions and derive some rates of convergence under differents sets of assumptions. Finally, we complete our theoretical results with a numerical study over simulated diffusion paths.

Keywords. Multiclass classification, Empirical risk minimization, Diffusion process, Margin assumption, Repeated observations.

Contents

4.1 Introduction

In this chapter, we tackle the multiclass supervised classification of trajectories of time-homogeneous diffusion processes. More precisely, we consider labelled trajectories generated by a random pair (X, Y) from a probability space $(\Omega, \mathcal{F}, \mathbb{P}_{X,Y})$, with the feature X coming from a mixture of diffusion processes, and the label $Y \in \mathcal{Y} = \{1, \ldots, K\}$, $K \geq 2$, a random variable of discrete law $\mathfrak{p}^* = (\mathfrak{p}_1^*, \ldots, \mathfrak{p}_K^*)$. We place ourselves in a nonparametric setting where for each label $i \in \mathcal{Y}$, and given $\{Y = i\}$, the diffusion process X is solution of a stochastic differential equation whose drift function b_i^* is assumed to be unknown and depends on the class i, and whose diffusion coefficient σ^*, also unknown, is common to all classes. Moreover, the discrete law $\mathfrak{p}^*$ of the label Y is also assumed to be unknown. We suppose to have at our disposal a learning sample $\mathcal{D}_N = \{(\bar{X}^j, Y_j),\ j = 1, \ldots, N\}$ composed of N independent copies of the random couple (X, Y), where the diffusion process X is observed at discrete times. We build from $\mathcal{D}_N$, an Empirical Risk Minimization (ERM) type classifier $\widehat{g}$ that mimics the Bayes classifier g^*, minimizer of the classification risk, that is,

$$g^* \in \arg\min_g \mathbb{P}_{X,Y}(g(X) \neq Y),$$

and whose performance is assessed by its excess risk $\mathbb{P}(\widehat{g}(X) \neq Y) - \mathbb{P}(g^*(X) \neq Y)$ with respect to the Bayes classifier.

The ERM type classifier is studied in the literature, for finite-dimensional data (see *e.g.* [71], [92]), and for functional data modelled by diffusion processes (see [11], [29]). Not far from our framework, the ERM-type classifiers proposed in [11] and [29] reach rates of convergence of order $N^{-\alpha/(2\alpha+1)}$ with α a regularity parameter. The rate is obtained in [11] in the context of binary classification, and in [29] in the multiclass classification setup, where the diffusion coefficient is assumed to be known. However, both ERM type classifiers are not implementable in practice. In this chapter, we propose an implementable multiclass classification procedure based on the convexification of the model (see [96], [8]), where the drift functions b_i^*, $i \in \mathcal{Y}$, and the diffusion coefficient σ^* are unknown.

The results in this chapter are resumed as follows:

i) We prove the consistency of the ERM type classifier, and derive a rate of convergence of order $N^{-\beta/(2\beta+1)}$ over the Hölder space $\Sigma(\beta, R)$ of regularity parameter $\beta \geq 1$, given by

$$\Sigma(\beta, R) = \left\{ f \in \mathcal{C}^{\lfloor\beta\rfloor+1}(\mathbb{R}) : \left| f^{(\ell)}(x) - f^{(\ell)}(y) \right| \leq R|x - y|,\ \ x, y \in \mathbb{R} \right\}$$

with $\ell = \lfloor \ell \rfloor$, and $R > 0$.

ii) In the binary classification setup, we make a margin assumption on the regression function associated to the diffusion model, and establish a rate of convergence of the resulting ERM type classifier of order $N^{-4\beta/3(2\beta+1)}$ over the Hölder space $\Sigma(\beta, R)$ with $\beta \geq 1$.

In Section 4.2, we present the diffusion model and the assumptions on the drift and diffusion coefficients. Section 4.3 is devoted to the construction of the nonparametric ERM type classifier. The theoretical properties of the empirical classifier are given in Section 4.4, and a faster rate of a binary ERM type classifier is established in Section 4.5. A numerical study over simulated data is made in Section 4.6, and Section 4.8 is devoted to the proofs of our results.

4.2 Model and assumptions

We consider a multiclass classification model in which the feature $X = (X_t)_{t\in[0,1]}$ is a diffusion process, solution of the following mixture model:

$$dX_t = b^*_Y(X_t)dt + \sigma^*(X_t)dW_t, \;\; t \in [0,1], \;\; X_0 = 0, \tag{4.1}$$

where $Y \in \mathcal{Y} = \{1,\dots,K\}$, $K \geq 2$, is a discrete random variable with an unknown distribution $\mathfrak{p}^* = (\mathfrak{p}^*_1,\dots,\mathfrak{p}^*_K)$, and $(W_t)_{t\geq 0}$ is a standard Brownian motion independent of Y. Thus, each label data is generated by the random couple (X,Y) where X is the feature and Y the corresponding class. We assume that the distribution of the random pair (X,Y) is unknown. More precisely, we assume that the coefficients b^*_i, $i \in \mathcal{Y}$, and σ^* of the diffusion processes are unknown. Moreover, we assume that $0 < \mathfrak{p}^*_0 = \min_{i\in\mathcal{Y}} \mathfrak{p}^*_i$, and we set $\mathbf{b}^* = (b^*_1,\dots,b^*_K)$.

The goal is to construct a measurable function g over the set of diffusion processes solutions of (4.1), and called classifier, whose role is to predict, for each feature X, a value $g(X) \in \mathcal{Y}$ which corresponds to the associated label Y. Let $\mathcal{G}$ be a set of classifiers $g : \mathcal{X} \longrightarrow \mathcal{Y}$ that will be defined in Section 4.3.2. The performance of each $g \in \mathcal{G}$ is assessed by $\mathcal{R}(g)$, where $\mathcal{R}$ is the classification risk defined in Chapter 1, Section 3.2.

Assumptions. We assume that the drift functions b^*_i, $i \in \mathcal{Y}$ and diffusion coefficient σ^* satisfy Assumptions 3.2.1 and 3.2.2 defined in Chapter 3. We recall that under Assumption 3.2.1, there exists a diffusion process X, unique strong solution of the model (4.1), which admits a transition density $p_X : (t,x) \mapsto p_X(t,x)$ given by

$$p_X = \sum_{i=1}^{K} \mathfrak{p}^*_i p_{i,X},$$

where each $p_{i,X}$ is the transition density of the diffusion process X given $\{Y = i\}$ (see *e.g.* [43], [52]).

Bayes classifier. We also recall that, under Assumptions 3.2.1 and 3.2.2, the Bayes classifier g^* defined in Chapter 3, Section 3.2.2 is characterized for all $X \in \mathcal{X}$ by

$$g^*(X) = \arg\max_{i\in\mathcal{Y}} \pi^*_i(X),$$

where,

$$\forall\, i \in \mathcal{Y}, \;\; \pi^*_i(X) = \phi^*_i\left(\mathbf{F}^*(X)\right), \quad \mathbf{F}^* = (F^*_1,\dots,F^*_k)$$

with, for each $i \in \mathcal{Y}$,

$$F^*_i(X) := \int_0^1 \frac{b^*_i}{\sigma^{*2}}(X_s)dX_s - \frac{1}{2}\int_0^1 \frac{b^{*2}_i}{\sigma^{*2}}(X_s)ds,$$

and $\phi^*_i : (x_1,\cdots,x_K) \mapsto \frac{\mathfrak{p}^*_i e^{x_i}}{\sum_{k=1}^K \mathfrak{p}^*_k e^{x_k}}$ are the softmax functions.

As we observe the diffusion paths in discrete time, denote by $\bar{g}^*$ the discrete time version of the Bayes classifier g^* defined by

$$\bar{g}^*(X) = \arg\max_{i\in\mathcal{Y}} \bar{\pi}^*_i(X),$$

where $\bar{\pi}^*_i(X) = \phi^*_i(\bar{\mathbf{F}}^*(X))$, with $\bar{\mathbf{F}}^* = \left(\bar{F}^*_1,\dots,\bar{F}^*_K\right)$, and for each $i \in \mathcal{Y}$,

$$\bar{F}^*_i = \sum_{k=0}^{n-1} \frac{b^*_i}{\sigma^{*2}}(X_{k\Delta_n})(X_{(k+1)\Delta_n} - X_{k\Delta_n}) - \frac{\Delta_n}{2}\sum_{k=0}^{n-1} \frac{b^{*2}_i}{\sigma^{*2}}(X_{k\Delta_n}).$$

4.3 Empirical classification procedure

Consider the learning sample $\mathcal{D}_N = \left\{(\bar{X}^j, Y_j),\ j = 1, \ldots, N\right\}$ composed of N independent copies of the random pair $(\bar{X}, Y)$ with $\bar{X} = (X_{k\Delta_n})_{0\leq k\leq n}$ with $\Delta_n = 1/n$ the time-step. The empirical classification risk $\widehat{\mathcal{R}} : \mathcal{G} \longrightarrow [0,1]$ is defined for all $g \in \mathcal{G}$ by

$$\widehat{\mathcal{R}}(g) := \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{g(\bar{X}^j)\neq Y_j},$$

and converges almost surely toward the theoretical risk $\mathcal{R}(g)$. The ERM type classifier $\widehat{g}$ is given by

$$\widehat{g} \in \arg\min_{g\in\mathcal{G}} \widehat{\mathcal{R}}(g). \tag{4.2}$$

In the literature, classification procedures of the ERM type have been proposed both for functional data (see [11], [29]), and for multivariate data (see *e.g.* [92], [71]).

4.3.1 Convexification of the problem

Since the set $\mathcal{G}$ of classifiers together with the empirical classification risk $\widehat{\mathcal{R}}$ are non-convex, the classifier $\widehat{g}$, solution of a non-convex optimization problem (4.2) is NP-hard to compute in practical situations, or even intractable considering the nature of the procedure (see *e.g.* [8], [3]). To overcome this problem and construct a classification procedure easier to compute in practice, we propose convex surrogates of both the loss function and the set $\mathcal{G}$ of classifiers as follows:

- The set of classifiers $\mathcal{G}$ is replaced by the convex space of score functions

  $$\mathcal{H} = \left\{\mathbf{h} = (h_1, \ldots, h_K) : \mathcal{X} \longrightarrow \mathbb{R}^K\right\}$$

  so that for each classifier $g \in \mathcal{G}$, there exists a score function $h \in \mathcal{H}$ such that for all $X \in \mathcal{X}$,

  $$g(X) = \arg\max_{i\in\mathcal{Y}} h^i(X).$$

  From now, we set $g = g_h$ to link a classifier $g \in \mathcal{G}$ to its corresponding score function $h \in \mathcal{H}$.

- The non-convex loss function $(y, y') \in \mathcal{Y}^2 \mapsto \mathbb{1}_{y\neq y'}$ is replaced by the quadratic loss function $x \mapsto (1-x)^2$. Then, the score functions h are now regarded as new classifiers whose performance is assessed by the new classification risk $\mathfrak{R} : \mathcal{H} \longrightarrow \mathbb{R}_+$ given for each $h \in \mathcal{H}$ by

  $$\mathfrak{R}(h) := \sum_{i=1}^{K} \mathbb{E}\left[(1 - Z_i h^i(X))^2\right]$$

  where $\mathbf{Z} = (Z_1, \ldots, Z_K)$ is a random vector with, for each $i \in \mathcal{Y}$, $Z_i = 2\mathbb{1}_{Y=i} - 1 \in \{-1, 1\}$. So the new Bayes classifier h^* is the minimizer of the classification risk $\mathfrak{R}$ over $\mathcal{H}$, and is characterized for all $X \in \mathcal{X}$ by

  $$h^{*i}(X) := 2\pi_i^*(X) - 1.$$

We define the empirical classifier $\widehat{h}$ by

$$\widehat{h} \in \arg\min_{h\in\mathcal{H}} \widehat{\mathfrak{R}}(h) \tag{4.3}$$

where $\widehat{\mathfrak{R}}$ is the new empirical risk of classification defined from the learning sample $\mathcal{D}_N$ and for each classifier $h \in \mathcal{H}$ by

$$\widehat{\mathfrak{R}}(h) := \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K} \left(1 - Z_i h^i(\bar{X}^j)\right)^2.$$

As a result, the empirical classifier $\widehat{g} = g_{\widehat{h}}$ is given for all $X \in \mathcal{X}$ by

$$\widehat{g}(X) = g_{\widehat{h}}(X) = \arg\max_{i \in \mathcal{Y}} \widehat{h}^i(X).$$

We establish the following result relating the excess risk $\mathcal{R}(g_{\widehat{h}}) - \mathcal{R}(g_{h^*})$ of the empirical classifier $g_{\widehat{h}}$ to the excess risk $\mathfrak{R}(\widehat{h}) - \mathfrak{R}(h^*)$ of the corresponding empirical classifier $\widehat{h}$.

Proposition 4.3.1 (Zhang (2004) [96]). *The empirical classifiers $\widehat{g} = g_{\widehat{h}}$ and $\widehat{h}$, respective solutions of Equations (4.3) and (4.4) , satisfy the following relation*

$$\mathcal{R}(g_{\widehat{h}}) - \mathcal{R}(g_{h^*}) \leq \frac{1}{\sqrt{2}} \sqrt{\mathfrak{R}(\widehat{h}) - \mathfrak{R}(h^*)}.$$

The result of Proposition 4.3.1 links the excess risk of any classifier $g = g_h \in \mathcal{G}$ to the excess risk of its corresponding classifier $h \in \mathcal{H}$. Thus, the consistency of the empirical classifier $\widehat{h}$ solution of the convex programming problem (4.3) implies, through Proposition 4.3.1 , the consistency of the corresponding empirical classifier $\widehat{g} = g_{\widehat{h}}$ solution of Equation (4.3). Furthermore, the result of Proposition 4.3.1 allows us to deduce the rate of convergence of the ERM type classifier $g_{\widehat{h}}$ from the rate of $\widehat{h}$.

4.3.2 The nonparametric ERM type classifier

Let $\mathcal{S}_{K_N}$ be the space of approximation of the unknown coefficients of the diffusion processes solution of (4.1), generated by the **B**-spline basis $\{B_\ell,\ \ell = -M, \ldots, K_N - 1\}$ built on the interval $[-\log(N), \log(N)]$, with $M, K_N \geq 1$, and given as follows:

$$\mathcal{S}_{K_N} := \left\{ \mathfrak{f} = \sum_{\ell=-M}^{K_N-1} a_\ell B_\ell,\ \sum_{\ell=-M}^{K_N-1} a_\ell^2 \leq (K_N + M)\log(N) \right\},$$

$$\widetilde{\mathcal{S}}_{K_N} := \left\{ x \mapsto \mathfrak{s}^2(x) = \mathfrak{f}(x)\mathbb{1}_{\mathfrak{f}(x) \geq 1/\log^2(N)} + \frac{1}{\log^2(N)}\mathbb{1}_{\mathfrak{f}(x) \leq 1/\log^2(N)},\ \ \mathfrak{f} \in \mathcal{S}_{K_N} \right\}.$$

Let $\mathcal{X}$ be the set of diffusion paths, and define the set of classifiers $\mathcal{G}$ by

$$\mathcal{G} = \left\{ g_{\mathbf{b},\sigma^2,\mathfrak{p}} : \mathcal{X} \longrightarrow \mathcal{Y} \mid\ (\mathbf{b}, \sigma^2) \in (\mathcal{S}_{K_N})^K \times \widetilde{\mathcal{S}}_{K_N},\ \ \mathfrak{p} \in (0,1)^K \right\},$$

where each $g_{\mathbf{b},\sigma^2,\mathfrak{p}}$ is a plug-in classifier of $g^* = g_{\mathbf{b}^*,\sigma^{*2},\mathbf{p}^*}$. Thus, the ERM type classifier $\widehat{g}$ is the minimizer of the empirical classification risk $\widehat{\mathcal{R}}$ on the set $\mathcal{G}$, that is,

$$\widehat{g} \in \arg\min_{g \in \mathcal{G}} \widehat{\mathcal{R}}(g). \tag{4.4}$$

Recall that the Bayes classifier h^* is given for all $X \in \mathcal{X}$ and for all $i \in \mathcal{Y}$ by

$$h^{*i}(X) = 2\pi_i^*(X) - 1.$$

Then, each coordinate function h^{*i} takes its values in the compact interval $[-1, 1]$. As a result, setting $\pi_i^* = \pi_{b_i^*,\sigma^{*2},\mathfrak{p}^*}$ for each $i \in \mathcal{Y}$, we restrict the convex space $\mathcal{H}$ to the space $\mathfrak{H}$ defined by

$$\mathfrak{H} := \left\{ h = (h^1, \ldots, h^K) \in \mathcal{H} \mid \forall\, i \in \mathcal{Y},\ h^i(X) = 2\pi_{\mathbf{b},\sigma^2,\mathfrak{p}}(X) - 1,\ \ \mathbf{b} \in (\mathcal{S}_{K_N}^b)^K,\ \ \sigma^2 \in \widetilde{\mathcal{S}}_{K_N},\ \mathfrak{p} \in (0,1)^K \right\}, \tag{4.5}$$

and we define the empirical classifier $\widehat{g} = g_{\widehat{h}}$, where $\widehat{h}$ is given by

$$\widehat{h} = \arg\min_{h \in \mathfrak{H}} \widehat{\mathfrak{R}}(h).$$

We rather consider in the sequel, the empirical classifier $\widehat{g}$ which has the benefit of being implementable thanks to the convexification of the problem.

4.4 Theoretical properties of the ERM type classifier

We study in this section the consistency of our ERM type classifier $\widehat{g}$ and derive a rate of convergence under mild assumptions on the unknown functions $\mathbf{b}^*$ and σ^{*2}. We define from the learning sample $\mathcal{D}_N$, the following two learning subsamples

$$\mathcal{D}_N^{(0)} = \left\{(\bar{X}^{0,j}, Y_j^0) \in \mathcal{D}_N, \;\; j = 1, \ldots, N_0\right\},$$

and

$$\mathcal{D}_N^{(1)} = \left\{(\bar{X}^{1,j}, Y_j^1) \in \mathcal{D}_N, \;\; j = 1, \ldots, N_1\right\}$$

of respective sizes N_0 and N_1, with $N_0 + N_1 = N$, and $N_0, N_1 \to \infty$ as $N \to \infty$ (for example $N_0 = \lfloor \frac{1}{4} N \rfloor$ and $N_1 = N - N_0$). Then, we build from the learning sample $\mathcal{D}_N^{(0)}$, the estimator $\widehat{\mathfrak{p}}$ of $\mathfrak{p}^*$ given by

$$\widehat{\mathfrak{p}}_i := \frac{1}{N_0} \sum_{j=1}^{N_0} \mathbb{1}_{Y_j^0 = i}, \;\; i \in \mathcal{Y},$$

and from the learning subsample $\mathcal{D}_N^{(1)}$, we build the empirical classifier $\widehat{g} = g_{\widehat{\mathbf{b}}, \widehat{\sigma}^2, \widehat{\mathfrak{p}}}$ as follows:

$$\widehat{g} \in \underset{(\mathbf{b}, \sigma^2) \in \mathfrak{W}_{K_N, L_N}}{\arg\min} \mathcal{R}(g_{\mathbf{b}, \sigma^2, \widehat{\mathfrak{p}}})$$

where $\mathfrak{W}_{K_N} = (\mathcal{S}_{K_N})^K \times \widetilde{\mathcal{S}}_{K_N}$. After the convexification of the model, we have that $\widehat{g} = g_{\widehat{h}}$ where the score function $\widehat{h}$ is built from the learning sample $\mathcal{D}_N^{(1)}$ by

$$\widehat{h} \in \underset{(\mathbf{b}, \sigma^2) \in \mathfrak{W}_{K_N}}{\arg\min} \widehat{\mathfrak{R}}(h_{\mathbf{b}, \sigma^2, \widehat{\mathfrak{p}}}).$$

We establish below an upper-bound of the average excess risk of the empirical classifier $\widehat{g}$.

Theorem 4.4.1. *Under Assumptions 3.2.1 and 3.2.2, the following holds,*

$$\mathbb{E}\left[\mathcal{R}(g_{\widehat{h}})\right] - \mathcal{R}(g_{h^*}) \leq C \left(\frac{\log^{3/2}(N)}{K_N} + \left(\frac{K_N \log(N)}{N} \right)^{1/2} + \Delta_n^{1/2} + \sqrt{\frac{n}{\log(N)}} \exp\left(-\frac{c}{2} \log^2(N) \right) \right)$$

where $C > 0$ is a constant independent of N and n.

We obtain the consistency of the ERM type classifier $\widehat{g} = g_{\widehat{h}}$ from the result of Theorem 4.4.1 with K_N of polynomial growth, and $n \propto N$. Note that the above result is obtained in the general case where les functions b_i^* and σ^* are unknown and non-constant.

If we assume that the functions b_i^* and σ^{*2} belong to the Hölder space $\Sigma(\beta, R)$, with $\beta \geq 1$ and $R > 0$, given by

$$\Sigma(\beta, R) := \left\{ f \in \mathcal{C}^{\lfloor \beta \rfloor + 1}(\mathbb{R}), \; \left| f^{(\ell)}(x) - f^{(\ell)}(y) \right| \leq R|x - y|^{\beta - \ell}, \; x, y \in \mathbb{R} \right\},$$

then we obtain the following result.

Corollary 4.4.2. *Suppose that $b_Y^*, \sigma^{*2} \in \Sigma(\beta, R)$, $K_N \propto N^{1/(2\beta+1)}$ and $N \propto n$. Under Assumptions 3.2.1 and 3.2.2, the following holds:*

$$\mathbb{E}\left[\mathcal{R}(g_{\widehat{h}})\right] - \mathcal{R}(g_{h^*}) = \mathrm{O}\left(N^{-\beta/(2\beta+1)}\right).$$

The obtained rate of convergence is of the same order than the one established in chapter 3 for the plug-in type classifier where the diffusion coefficient is known and constant. Note that, contrary to the case of the plug-in type classifier which is based on the nonparametric estimators of the coefficients of the diffusion processes, the ERM type classifier focuses on the minimization of the empirical

classification risk which approximated the performance measurement for classifiers. In a nonparametric multiclass classification framework, and in a non-convex setup, Denis *et al.* (2020) ([29]) built a ERM type classifier, where the diffusion coefficient is assumed to be known. They established a rate of order $N^{-\beta/(2\beta+1)}$ over a Hölder space of smoothness parameter $\beta \geq 1$. In our framework, we consider the set of classifiers

$$\mathcal{G} = \left\{ g_{\mathbf{b}},\ \mathbf{b} = (b_1, \ldots, b_K) \in \mathcal{S}_{K_N}^K \right\},$$

and we derive the following result.

Theorem 4.4.3. *Suppose that $b_i^*, \sigma^{*2} \in \Sigma(\beta, R)$ with $\beta \geq 1$ and $R > 0$. Under Assumptions 3.2.1 and 3.2.2, the following holds:*

$$\mathbb{E}\left[\mathcal{R}(\widehat{g})\right] - \mathcal{R}(g^*) = \mathrm{O}\left(N^{-\beta/(2\beta+1)}\right).$$

The result of Theorem 4.4.3 is obtained using the similar arguments as in Denis *et al.* (see [29]). In the next section, we study a faster rate of convergence under stronger assumptions on our model.

4.5 Faster rate of convergence under margin assumption

We suppose that $K = 2$, $\mathfrak{p}^* = (1/2, 1/2)$ the known discrete law of the label $Y = \{-1, 1\}$ and the diffusion coefficient σ^* is known with $\sigma^* = 1$. Denote by η^* the regression function given for each feature X by $\eta^*(X) = \mathbb{P}(Y = 1|X)$. Let $\widetilde{\mathbb{P}}$ be a probability measure under which the diffusion process $X = (X_t)_{t\in[0,1]}$ is solution of the S.D.E. $dX_t = \sigma(X_t)d\widetilde{W}_t$, where $\widetilde{W}$ is a brownian motion under $\widetilde{\mathbb{P}}$. Under Assumptions 3.2.1 and 3.2.2, the Girsanov's theorem (see [81]) implies

$$\frac{d\mathbb{P}_0}{d\widetilde{\mathbb{P}}}(X) = \exp\left(\int_0^1 \frac{b_0^*}{\sigma^2}(X_s)dX_s - \int_0^1 \frac{b_0^{*2}}{2\sigma^2}(X_s)ds\right) =: q_0(X)$$
$$\frac{d\mathbb{P}_1}{d\widetilde{\mathbb{P}}}(X) = \exp\left(\int_0^1 \frac{b_1^*}{\sigma^2}(X_s)dX_s - \int_0^1 \frac{b_1^{*2}}{2\sigma^2}(X_s)ds\right) =: q_1(X)$$

where $\mathbb{P} = \frac{1}{2}\mathbb{P}_0 + \frac{1}{2}\mathbb{P}_1$. Then we have

$$\eta^*(X) := \mathbb{P}(Y = 1|X) = \frac{\mathbb{P}(Y = 1)\frac{d\mathbb{P}_1}{d\widetilde{P}}(X)}{\mathbb{P}(Y = 0)\frac{d\mathbb{P}_0}{d\widetilde{\mathbb{P}}}(X) + \mathbb{P}(Y = 1)\frac{d\mathbb{P}_1}{d\widetilde{\mathbb{P}}}(X)} = \frac{\frac{d\mathbb{P}_1}{d\widetilde{\mathbb{P}}}(X)}{\frac{d\mathbb{P}_0}{d\widetilde{\mathbb{P}}}(X) + \frac{d\mathbb{P}_1}{d\widetilde{\mathbb{P}}}(X)}.$$

4.5.1 The Bayes Classifier

In binary classification, the Bayes classifier g^* is defined from the regression function η^* by $g^* = \mathbb{1}_{\eta^*\geq 1/2}$. Thus, for any diffusion process X, we have

$$g^*(X) = \mathbb{1}_{\eta(X)\geq\frac{1}{2}} = \mathbb{1}_{\frac{d\mathbb{P}_1}{d\widetilde{\mathbb{P}}}(X)\geq\frac{d\mathbb{P}_0}{d\widetilde{\mathbb{P}}}(X)} = \mathbb{1}_{\int_0^1 \frac{b_1^*-b_0^*}{\sigma^2}(X_s)dX_s \geq \int_0^1 \frac{b_1^{*2}-b_0^{*2}}{2\sigma^2}(X_s)ds}.$$

As we already know, the Bayes classifier g^* minimizes the classification error $g \mapsto \mathcal{R}(g) = \mathbb{P}(g(X) \neq Y)$. As a result, the performance of any classifier $g \in \mathcal{G}$ is assessed through its excess risk with respect to g^* given as follows:

$$\mathcal{R}(g) - \mathcal{R}(g^*) = \mathbb{E}\left[|2\eta^*(X) - 1|\,\mathbb{1}_{g(X)\neq g^*(X)}\right],$$

(see *e.g.* [44], [29]).

4.5.2 Control of the margin

We make the following assumptions on the drift functions b_0^* and b_1^*.

Assumption 4.5.1. *The drift functions b_0^* and b_1^* are bounded on the real line, and the random variable*

$$Z := \frac{1}{\Delta_{b^*}} \int_0^1 (b_1^* - b_0^*)(X_s) dW_s, \text{ with } \Delta_{b^*}^2 = \mathbb{E}\left[\int_0^1 (b_1^* - b_0^*)^2(X_s) ds\right] > 0,$$

has a density function that is bounded on $\mathbb{R}$.

The existence of a density function of the random variable Z is proven using Maliavin calculus and the Hörmander's condition (see [74], Chapter 2, Theorem 2.3.3). The strictly positive constant $\underline{\Delta}_b$ which is the minimum separation distance between the respective curves of the drift functions b_0^* and b_1^* characterizes the margin between the two classes. The condition $\underline{\Delta}_b > 0$ implies that the two classes do not mix. We obtain the following result.

Proposition 4.5.2. *Under Assumptions 3.2.1 and 4.5.1 and for all $\varepsilon \in (0, 1/8)$, there exists a constant $C > 0$ such that*

$$\mathbb{P}\left(0 < \left|\eta^*(X) - \frac{1}{2}\right| \leq \varepsilon\right) \leq C \frac{12}{\Delta_{b^*}} \varepsilon.$$

We deduce from [8], *Lemma 5 (and its proof)* that the result of Proposition 4.5.2 implies that for all classifier $g \in \mathcal{G}$,

$$\mathbb{P}\left(g(X)(\eta^*(X) - 1/2) < 0\right) \leq \frac{c}{\sqrt{\Delta_{b^*}}} (\mathcal{R}(g) - \mathcal{R}(g^*))^{1/2}$$

where $c > 0$ is a constant. Moreover, *Theorem 3* in [8] implies that for all classifier $g \in \mathcal{G}$, there exist a classifier $h \in \mathfrak{H}$ and a constant $c > 0$ such that

$$\mathcal{R}(g) - \mathcal{R}(g^*) \leq \frac{c}{\Delta_{b^*}^{1/3}} \left(\mathfrak{R}(h) - \mathfrak{R}(h^*)\right)^{2/3}. \tag{4.6}$$

We finally deduce from Theorem 4.4.1, Assumption 4.5.1 and Equation (4.6) the following result.

Theorem 4.5.3. *Suppose that $b_0^*, b_1^* \in \Sigma(\beta, R)$. Under Assumptions 3.2.1, 3.2.2 and 4.5.1, the following holds:*

$$\mathbb{E}\left[\mathcal{R}(g_{\widehat{h}})\right] - \mathcal{R}(g_{h^*}) = \mathrm{O}\left(N^{-4\beta/3(2\beta+1)}\right).$$

where $g_{\widehat{h}} = \widehat{g}$ and $g_{h^*} = g^*$.

Thus, we have a faster rate of convergence compared to the result of Theorem 4.4.2 thanks to the margin assumption characterized by Proposition 4.5.2 which minimizes the wrong prediction of labels when the probability of belonging of the new observation to the class 1 is too close to $1/2$.

4.6 Numerical illustration

In this section, we make a numerical study of the ERM type classifier over synthetic data. We present the diffusion models in Section 4.6.1, and the numerical results in Section 4.6.2.

4.6.1 Models and simulations

We fix the number of classes to $K = 3$, and the discrete law of the label $Y \in \mathcal{Y} = \{1, \ldots, 3\}$ is $\mathfrak{p}^* = \{1/3, 1/3, 1/3\}$. We study the two following diffusion models:

Model 1: $b(\theta, x) = -(x - \theta)$, $\theta \in \{-1, 0, 1\}$, $\sigma(x) = 1$

Model 2: $b(\theta, x) = \theta\left[1/4 + (3/4)\cos^2 x\right]$, $\theta \in \{1, 3/2, -3/2\}$, $\sigma(x) = 0.1 + 0.9/\sqrt{1 + x^2}$.

	Model 1	Model 2
$\widehat{\mathcal{R}}(g^*)$	0.41 (0.01)	0.36 (0.01)

Table 4.1: *Classification risks of the Bayes classifier g^* from learning samples of size $N = 4000$ with $n = 500$.*

The two diffusion models satisfy Assumptions 3.2.1 and 3.2.2. We investigate the numerical consistency of the empirical classifier $\widehat{g} = g_{\widehat{h}}$ using learning samples of size $N \in \{100, 1000\}$ composed of trajectories of length $n = 100$. We use a learning sample of size $N = 4000$ with diffusion paths of length $n = 500$ for the evaluation of the Bayes classifier g^*.

We deduce from the result of Table 4.1 that Model 2 is simpler than Model 1 since the Bayes classifier of Model 2 has a better performance. Moreover, from the definition of the Bayes classifier, we known that the ERM type classifier could have at most the performance of the Bayes classifier given in Table 4.1.

4.6.2 Numerical results

We build the ERM type classifier from learning samples of size $N \in \{100, 1000\}$, with each diffusion path of length $n = 100$. The approximation subspace $\mathcal{S}_{K_N}$ spanned by the **B**-spline basis is chosen such that $M = 3$, $K_N = 4$ for the approximation of the drift function and of the square of the diffusion coefficient. The **B**-spline basis is built on the compact interval $[-A_N, A_N]$ with $A_N = 10$. We evaluate the performance of the empirical classifier on test samples of size $N_{\text{test}} = 10000$ over 20 repetitions. The table below compares the performance of our classifier with that of the classifiers for functional data based on the random forest and the notion of depth for functional data.

$\widehat{\mathcal{R}}(\widehat{g})$	ERM		Random Forest		Depth	
	$N = 100$	$N = 1000$	$N = 100$	$N = 1000$	$N = 100$	$N = 1000$
Model 1	0.43 (0.01)	0.41 (0.01)	0.46 (0.02)	0.44 (0.01)	0.44 (0.02)	0.43 (0.01)
Model 2	0.38 (0.02)	0.36 (0.01)	0.40 (0.01)	0.38 (0.01)	0.40 (0.02)	0.39 (0.01)

Table 4.2: *Performance of the ERM type classifier, and comparison with that of the random forest and the depth based classifier for diffusion paths collected at discrete times with the grid $n = 100$.*

We observe from Table 4.2 that the ERM type classifier performs better with respect to the classifiers for functional data based respectively on the random forest and the depth of functional data. The main benefit with our classifier is that its construction takes into account the theoretical properties of the diffusion processes that generate the data.

The diffusion models used for this illustration are exactly the same chosen for the numerical study of the performance of the plug-in type classifier studied in Chapter 3. We remark that the ERM type classifier performs better compared to the plug-in type classifier. This result is explained by the fact that contrary to the plug-in type classifier, the construction of the ERM type classifier is based on the minimization of the empirical risk of classification. Note that the evaluation of the numerical performance of the ERM type classifier is not optimal since there is not a data-driven selection of dimension of the approximation space.

4.7 Conclusion and discussion

In this chapter, we have built a nonparametric classification procedure by minimizing the empirical risk of classification from the learning sample. We placed ourselves in the case where the drift functions $b_i^*,\ i \in \mathcal{Y}$ and the diffusion coefficient σ^* are unknown, and we derived, under mild assumptions on these unknown functions, a rate of convergence of the resulting ERM type classifier of

order $N^{-\beta/(2\beta+1)}$ over the Hölder space of smoothness parameter $\beta \geq 1$. In addition to the fact that our empirical classifier is implementable thanks to the convexification of the model, this theoretical result extends the results established in [11] and [29] to a nonparametric setup with an unknown diffusion coefficient. Finally, under a margin assumption on the regression function in binary classification setup, we derived a faster rate of order $N^{-4\beta/3(2\beta+1)}$ over the Hölder space $\Sigma(\beta, R)$ with $\beta \geq 1$.

It should be noted, however, that the nonparametric classification procedure built in this chapter is based on the minimization of a convexified optimization problem over a finite-dimensional approximation space whose dimension is not optimally selected from the data. As a result, a immediate perspective could be to propose an adaptive ERM type classifier based on a data-driven selection of the dimension of the approximation space. Moreover, one may also focus on the study of the optimality of the obtained rates of convergence both under mild assumptions on the coefficients of the diffusion processes and under a margin assumption on the regression function associated to the model. We could also extend the obtained results to the case of in-homogeneous diffusion processes.

4.8 Proofs

For simplicity, we denote the Bayes classifier g_{h^*} by g^*, and any other classifier g_h by g, where $h \in \mathfrak{H}$ is a score function.

4.8.1 Proof of Theorem 4.4.1

Proof. According to the result of Proposition 4.3.1

$$\mathbb{E}\left[\mathcal{R}(\widehat{g})\right] - \mathcal{R}(g^*) \leq \sqrt{\mathbb{E}\left[\mathfrak{R}(\widehat{h})\right] - \mathfrak{R}(h^*)}.$$

Denote by $\bar{h}_*$, the discrete time version of the Bayes classifier h^*. We have the following decomposition of the excess risk of $\widehat{h}$,

$$\mathfrak{R}(\widehat{h}) - \mathfrak{R}(h^*) = \mathfrak{R}(\widehat{h}) - \mathfrak{R}(\bar{h}_*) + \mathfrak{R}(\bar{h}_*) - \mathfrak{R}(h^*)$$

Upper bound of $\mathfrak{R}(\bar{h}_*) - \mathfrak{R}(h^*)$

$$\begin{aligned}
\mathfrak{R}\left(\bar{h}_*\right) - \mathfrak{R}(h^*) &= 4\sum_{i=1}^{K} \mathbb{E}\left|\bar{\pi}_i^*(X) - \pi_i^*(X)\right|^2 \\
&= 4\sum_{i=1}^{K} \mathbb{E}\left|\ \phi_i^*\left(\bar{\mathbf{F}}^*(X)\right) - \phi_i^*\left(\mathbf{F}^*(X)\right)\right|^2 \\
&\leq 4K\sum_{i=1}^{K} \mathbb{E}\left|\bar{F}_i^*(X) - F_i^*(X)\right|^2.
\end{aligned}$$

We set $\xi(s) = k\Delta$ if $s \in [k\Delta, (k+1)\Delta[$. For all $i \in \mathcal{Y}$,

$$\begin{aligned}
F_i^*(X) - \bar{F}_i^*(X) = &\int_0^1 \left(\frac{b_i^*}{\sigma^{*2}}(X_{\xi(s)}) - \frac{b_i^*}{\sigma^{*2}}(X_s)\right) b_Y^*(X_s)ds \\
&- \frac{1}{2}\int_0^1 \left(\frac{b_i^{*2}}{\sigma^{*2}}(X_{\xi(s)}) - \frac{b_i^{*2}}{\sigma^{*2}}(X_s)\right) ds \\
&+ \int_0^1 \left(\frac{b_i^*}{\sigma^{*2}}(X_{\xi(s)}) - \frac{b_i^*}{\sigma^{*2}}(X_s)\right) \sigma(X_s)dW_s.
\end{aligned}$$

Then we have:

$$\mathbb{E}\left|F_i^*(X)-\bar{F}_i^*(X)\right|^2 \leq 3\mathbb{E}\left|\int_0^1\left(\frac{b_i^*}{\sigma^{*2}}(X_{\xi(s)})-\frac{b_i^*}{\sigma^{*2}}(X_s)\right)b_Y^*(X_s)ds\right|^2 + \frac{3}{4}\mathbb{E}\left|\int_0^1\left(\frac{b_i^{*2}}{\sigma^{*2}}(X_{\xi(s)})-\frac{b_i^{*2}}{\sigma^{*2}}(X_s)\right)ds\right|^2 + 3\mathbb{E}\left|\int_0^1\left(\frac{b_i^*}{\sigma^{*2}}(X_{\xi(s)})-\frac{b_i^*}{\sigma^{*2}}(X_s)\right)\sigma^*(X_s)dW_s\right|^2.$$

According to the proof of Theorem 5.1 or Theorem 3.3 in [28], we obtain:

$$\forall i\in\mathcal{Y},\ \ \mathbb{E}\left|F_i^*(X)-\bar{F}_i^*(X)\right|^2\leq C\Delta$$

where $C>0$ is a constant. Finally, there exists a constant $C>0$ such that

$$\mathfrak{R}\left(\bar{h}_*\right)-\mathfrak{R}(h^*)\leq C\Delta.$$

Thus, the excess risk of the empirical classifier $\widehat{h}$ satisfies:

$$\mathbb{E}\left[\mathfrak{R}(\widehat{h})\right]-\mathfrak{R}(h^*)\leq\mathbb{E}\left[\mathfrak{R}(\widehat{h})\right]-\mathfrak{R}(\bar{h}_*)+C\Delta \tag{4.7}$$

where $C>0$ is a constant.

Upper bound of $\mathbb{E}\left[\mathfrak{R}(\widehat{h})\right]-\mathfrak{R}(\bar{h}_*)$

Let $\widetilde{h}$ be the approximation of the discrete version $\bar{h}_*$ of the Bayes classifier h^* given by

$$\widetilde{h}^i(X)=2\widetilde{\pi}_i(X)-1,\quad i\in[\![1,K]\!]\text{ and }X\in\mathcal{X}$$

where $\widetilde{\pi}_i$ depends on the functions $\widetilde{b}=\left(\widetilde{b}_1,\ldots,\widetilde{b}_K\right)$ and $\widetilde{\sigma}^2$, the respective approximations of $b^*=(b_1^*,\ldots,b_K^*)$ and σ^{*2}, with $\widetilde{b}_i\in\mathcal{S}_{K_N}$ and $\widetilde{\sigma}^2\in\widetilde{\mathcal{S}}_{K_N}$, such that $\|\widetilde{b}_i\|_\infty\leq\log^{3/2}(N),\|\widetilde{\sigma}^2\|_\infty\leq\log^{3/2}(N)$, and for all $x\in(-\log(N),\log(N))$,

$$\left|\widetilde{b}_i(x)-b_i^*(x)\right|\leq C\frac{\log(N)}{K_N},\text{ and }\left|\widetilde{\sigma}^2(x)-\sigma^{*2}(x)\right|\leq C\frac{\log(N)}{K_N}, \tag{4.8}$$

where $C>0$ is a constant (see [28]). Moreover, define

$$h^0=\left(h^{0,1},\ldots,h^{0,K}\right)=\arg\min_{h\in\mathfrak{H}}\mathfrak{R}(h),$$

where the space $\mathfrak{H}$ is given by Equation (4.5). For all classifier $h\in\mathfrak{H}$, set $\mathcal{D}_h:=\mathfrak{R}(h)-\mathfrak{R}(h^0)$ and $\widehat{\mathcal{D}}_h:=\widehat{\mathfrak{R}}(h)-\widehat{\mathfrak{R}}(h^0)$. One obtains that:

$$\begin{aligned}\mathfrak{R}(\widehat{h})-\mathfrak{R}(\bar{h}_*)&=\mathfrak{R}(\widehat{h})-\mathfrak{R}(h^0)+\mathfrak{R}(h^0)-\mathfrak{R}(\widetilde{h})+\mathfrak{R}(\widetilde{h})-\mathfrak{R}(\bar{h}_*)\\&=\mathcal{D}_{\widehat{h}}+\mathfrak{R}(h^0)-\mathfrak{R}(\widetilde{h})+\mathfrak{R}(\widetilde{h})-\mathfrak{R}(\bar{h}_*)\\&\leq\mathcal{D}_{\widehat{h}}-2\widehat{\mathcal{D}}_{\widehat{h}}+\mathfrak{R}(h^0)-\mathfrak{R}(\widetilde{h})+\mathfrak{R}(\widetilde{h})-\mathfrak{R}(\bar{h}_*),\end{aligned}$$

where $\widehat{\mathcal{D}}_{\widehat{h}}\leq 0$. We have $\mathfrak{R}(h^0)-\mathfrak{R}(\widetilde{h})\leq 0$ by definition of h^0, then

$$\mathfrak{R}(\widehat{h})-\mathfrak{R}(\bar{h}_*)\leq\mathcal{D}_{\widehat{h}}-2\widehat{\mathcal{D}}_{\widehat{h}}+\mathfrak{R}(\widetilde{h})-\mathfrak{R}(\bar{h}_*) \tag{4.9}$$

$$\mathfrak{R}(\widetilde{h})-\mathfrak{R}(\bar{h}_*)=4\sum_{i=1}^K\mathbb{E}\left[\left|\widetilde{\pi}_i(X)-\bar{\pi}_i^*(X)\right|^2\right]\leq 4K\sum_{i=1}^K\mathbb{E}\left[\left|\widetilde{F}_i(X)-\bar{F}_i^*(X)\right|^2\right].$$

For each $i \in [\![1, K]\!]$,

$$\mathbb{E}\left[\left|\widetilde{F}_i(X) - \bar{F}_i^*(X)\right|^2\right] \leq 3\mathbb{E}\left[\left|\int_0^1 \left(\frac{\widetilde{b}_i}{\widetilde{\sigma}^2}(X_{\xi(s)}) - \frac{b_i^*}{\sigma^{*2}}(X_{\xi(s)})\right) b_Y^*(X_s)ds\right|^2\right] + \frac{3}{4}\mathbb{E}\left[\left|\int_0^1 \left(\frac{\widetilde{b}_i^2}{\widetilde{\sigma}^2}(X_{\xi(s)}) - \frac{b_i^{*2}}{\sigma^{*2}}(X_{\xi(s)})\right) ds\right|^2\right] + 3\mathbb{E}\left[\left|\int_0^1 \left(\frac{\widetilde{b}_i}{\widetilde{\sigma}^2}(X_{\xi(s)}) - \frac{b_i^*}{\sigma^{*2}}(X_{\xi(s)})\right) \sigma^*(X_s)dW_s\right|^2\right].$$

Since for all $i \in [\![1, K]\!]$ and for all $x \in \mathbb{R}$, we have

$$\begin{cases} \left|\frac{\widetilde{b}_i}{\widetilde{\sigma}^2}(x) - \frac{b_i^*}{\sigma^{*2}}(x)\right| \leq \log^2(N)\left|\widetilde{b}_i(x) - b_i^*(x)\right| + \frac{\log^2(N)}{\sigma_0^{*2}}\left|b_i^*(x)\right|\left|\widetilde{\sigma}^2(x) - \sigma^{*2}(x)\right|, \\ \left|\frac{\widetilde{b}_i^2}{\widetilde{\sigma}^2}(x) - \frac{b_i^{*2}}{\sigma^{*2}}(x)\right| \leq \log^2(N)\left|\widetilde{b}_i(x) + b_i^*(x)\right|\left|\widetilde{b}_i(x) - b_i^*(x)\right| + \frac{\log^2(N)}{\sigma_0^{*2}}\left|b_i^*(x)\right|^2\left|\widetilde{\sigma}^2(x) - \sigma^{*2}(x)\right|. \end{cases}$$

Then, from Equation (4.8), there exists a constant $C > 0$ such that for all $x \in (-\log(N), \log(N))$,

$$\begin{cases} \left|\frac{\widetilde{b}_i}{\widetilde{\sigma}^2}(x) - \frac{b_i^*}{\sigma^{*2}}(x)\right| \leq C(1 + b_i^*(x))\frac{\log(N)}{K_N}, \\ \left|\frac{\widetilde{b}_i^2}{\widetilde{\sigma}^2}(x) - \frac{b_i^{*2}}{\sigma^{*2}}(x)\right| \leq C\left(\sqrt{L_N} + |b_i^*(x)| + |b_i^*(x)|^2\right)\frac{\log(N)}{K_N}. \end{cases}$$

Then, under Assumption 3.2.1, we have $\sup_{t\in[0,1]} \mathbb{E}[|b_Y^*(X_t)|^p] < \infty$ for all $p \in \mathbb{N}$, and there exists a constant $C > 0$ such that

$$\mathbb{E}\left[\left|\widetilde{F}_i(X) - \bar{F}_i^*(X)\right|^2 \mathbb{1}_{\left\{\max_{k=1,\dots,n}|X_{k\Delta}| < \log(N)\right\}}\right] \leq C\frac{\log^3(N)}{K_N^2}. \tag{4.10}$$

Moreover, under Assumption 3.2.1, we have

$$\begin{aligned} \mathbb{E}\left[\left|\widetilde{F}_i(X) - \bar{F}_i^*(X)\right|^2 \mathbb{1}_{\left\{\max_{k=1,\dots,n}|X_{k\Delta}| \geq \log(N)\right\}}\right] &\leq C\mathbb{P}\left(\max_{k=1,\dots,n}|X_{k\Delta}| \geq \log(N)\right) \\ &\leq C\sum_{k=1}^n \mathbb{P}\left(|X_{k\Delta}| \geq \log(N)\right) \\ &\leq Cn\sup_{t\in[0,1]}\mathbb{P}\left(|X_t| \geq \log(N)\right). \end{aligned}$$

From [28], *Lemma 7.3*, there exist constants $C, c > 0$ such that

$$\mathbb{E}\left[\left|\widetilde{F}_i(X) - \bar{F}_i^*(X)\right|^2 \mathbb{1}_{\left\{\max_{k=1,\dots,n}|X_{k\Delta}| \geq \log(N)\right\}}\right] \leq C\frac{n}{\log(N)}\exp\left(-c\log^2(N)\right). \tag{4.11}$$

Finally, we deduce from Equations (4.10) and (4.11) that

$$\mathfrak{R}(\widetilde{h}) - \mathfrak{R}(\bar{h}_*) \leq C\left(\frac{\log^3(N)}{K_N^2} + \frac{n}{\log(N)}\exp\left(-c\log^2(N)\right)\right). \tag{4.12}$$

Then we obtain from Equation (4.9) and (4.12) that

$$\mathbb{E}\left[\mathfrak{R}(\widehat{h})\right] - \mathfrak{R}(\bar{h}_*) \leq \mathbb{E}\left[\mathcal{D}_{\widehat{h}} - 2\widehat{\mathcal{D}}_{\widehat{h}}\right] + C\left(\frac{\log^3(N)}{K_N^2} + \frac{n}{\log(N)}\exp\left(-c\log^2(N)\right)\right). \tag{4.13}$$

We recall that the score function $\widehat{h}$ depends on functions $\widehat{b}_i,\ i \in \mathcal{Y}$ and $\widehat{\sigma}^2$ of respective coordinate vectors $\widehat{a^{(i)}},\ i \in \mathcal{Y}$ and $\widehat{\alpha}$. $\widehat{\mathbf{a}} = \left(\widehat{a^{(1)}}, \cdots, \widehat{a^{(K)}}\right)$. The coordinate vectors $\widehat{a^{(1)}}, \cdots, \widehat{a^{(K)}}, \widehat{\alpha}$ belong to the closed ball

$$\bar{B}' := \left\{\theta \in \mathbb{R}^{K_N+M},\ \|\theta\|_2 \leq \sqrt{(K_N+M)L_N}\right\} = \bar{B}_{\|.\|_2}\left(0, \sqrt{(K_N+M)L_N}\right),$$

and $\widehat{\mathfrak{p}}$ belong to the unit ball $\bar{B} = \bar{B}_{\|.\|_2}(0,1)$. Let $\bar{B}'_\varepsilon$ and $\bar{B}_\varepsilon$ be the respective $\varepsilon-$net of $\bar{B}'$ and $\bar{B}$ for all $\varepsilon > 0$. Let $\widehat{a^{(1)}}_\varepsilon, \cdots, \widehat{a^{(K)}}_\varepsilon, \widehat{\alpha}_\varepsilon \in \bar{B}'_\varepsilon$ and $\widehat{\mathfrak{p}}_\varepsilon \in \bar{B}_\varepsilon$ such that:

$$\left\|\widehat{a^{(i)}}_\varepsilon - \widehat{a^{(i)}}\right\|_2 \leq \varepsilon,\ i \in [\![1, K]\!],\ \ \|\widehat{\alpha}_\varepsilon - \widehat{\alpha}\|_2 \leq \varepsilon,\ \ \left\|\widehat{\mathfrak{p}}_\varepsilon - \widehat{\mathfrak{p}}\right\|_2 \leq \varepsilon.$$

Set $\widehat{h}_\varepsilon = \bar{h}_{\widehat{a}_\varepsilon, \widehat{\alpha}_\varepsilon, \widehat{p}_\varepsilon}$ with $\widehat{\mathbf{a}}_\varepsilon = \left(\widehat{a^1}_\varepsilon, \cdots, \widehat{a^K}_\varepsilon\right)$. We have:

$$\mathcal{D}_{\widehat{h}} - 2\widehat{\mathcal{D}}_{\widehat{h}} = \mathcal{D}_{\widehat{h}} - \mathcal{D}_{\widehat{h}_\varepsilon} + 2\left(\widehat{\mathcal{D}}_{\widehat{h}_\varepsilon} - \widehat{\mathcal{D}}_{\widehat{h}}\right) + \mathcal{D}_{\widehat{h}_\varepsilon} - 2\widehat{\mathcal{D}}_{\widehat{h}_\varepsilon}.$$

It results that,

$$\mathbb{E}\left[\mathcal{D}_{\widehat{h}} - 2\widehat{\mathcal{D}}_{\widehat{h}}\right] \leq \mathbb{E}\left(\mathcal{D}_{\widehat{h}} - \mathcal{D}_{\widehat{h}_\varepsilon}\right) + 2\mathbb{E}\left(\widehat{\mathcal{D}}_{\widehat{h}_\varepsilon} - \widehat{\mathcal{D}}_{\widehat{h}}\right) + \mathbb{E}\left(\mathcal{D}_{\widehat{h}_\varepsilon} - 2\widehat{\mathcal{D}}_{\widehat{h}_\varepsilon}\right). \tag{4.14}$$

1. Upper bound of $\mathbb{E}\left(\mathcal{D}_{\widehat{h}} - \mathcal{D}_{\widehat{h}_\varepsilon}\right)$

   Set $\widehat{b}_{i,\varepsilon} = \sum_{\ell=-M}^{K_N-1} \widehat{a^i}_\varepsilon B_\ell$ for all label $i \in [\![1, K]\!]$, $\widehat{b}_\varepsilon = \left(\widehat{b}_{1,\varepsilon}, \cdots, \widehat{b}_{K,\varepsilon}\right)$, and $\widehat{\sigma}^2_\varepsilon = \sum_{\ell=-M}^{K_N-1} \widehat{\alpha}_\varepsilon B_\ell$. We have:

$$\begin{aligned}\mathcal{D}_{\widehat{h}} - \mathcal{D}_{\widehat{h}_\varepsilon} = \mathfrak{R}(\widehat{h}) - \mathfrak{R}\left(\widehat{h}_\varepsilon\right) =\ & 4\sum_{i=1}^{K} \mathbb{E}\left|\widehat{\pi}_i(X) - \widehat{\pi}_i^\varepsilon(X)\right|^2 \leq C\sum_{i=1}^{K} \mathbb{E}\left|\widehat{F}_i(X) - \widehat{F}_i^\varepsilon(X)\right|^2 \\ & \leq C\sum_{i=1}^{K}\left\|\widehat{b}_i - \widehat{b}_{i,\varepsilon}\right\|_n^2 + \sum_{i=1}^{K}\left\|\widehat{b}_i\left(\widehat{\sigma}^2 - \widehat{\sigma}^2_\varepsilon\right)\right\|_n^2\end{aligned}$$

   according to [28], *Proof of Theorem 3.3*. For all function $g = \sum_{\ell=-M}^{K_N-1} a_\ell B_\ell$ belonging to any of the spaces $\mathcal{S}_{K_N}$ and $\widetilde{\mathcal{S}}_{K_N}$, using the Cauchy Schwartz inequality, we have:

$$\forall x \in \mathbb{R},\ \ g^2(x) = \left(\sum_{\ell=-M}^{K_N-1} a_\ell B_\ell(x)\right)^2 \leq \|a\|_2^2 \leq (K_N+M)\log^3(N).$$

   It results that for all g in $\mathcal{S}_{K_N}$ or $\widetilde{\mathcal{S}}_{K_N}$, $\|g\|_\infty \leq \|a\|_2 \leq \sqrt{(K_N+M)\log^3(N)}$ and,

$$\mathcal{D}_{\widehat{h}} - \mathcal{D}_{\widehat{h}_\varepsilon} \leq\ C\left(\sum_{i=1}^{K}\left\|\widehat{b}_i - \widehat{b}_{i,\varepsilon}\right\|_n^2 + (K_N+M)L_N\sum_{i=1}^{K}\left\|\widehat{\sigma}^2 - \widehat{\sigma}^2_\varepsilon\right\|_n^2\right)$$

   where the empirical norm $\|.\|_n$ is given for all g by

$$\|g\|_n := \sqrt{\mathbb{E}\left[\frac{1}{n}\sum_{k=0}^{n-1} g^2(X_{k\Delta})\right]} \leq \|g\|_\infty.$$

   Finally, there exists a constant $C > 0$ such that

$$\begin{aligned}\mathcal{D}_{\widehat{h}} - \mathcal{D}_{\widehat{h}_\varepsilon} &\leq C\left(\sum_{i=1}^{K}\left\|\widehat{b}_i - \widehat{b}_{i,\varepsilon}\right\|_\infty^2 + K_N L_N\left\|\widehat{\sigma}^2 - \widehat{\sigma}^2_\varepsilon\right\|_\infty^2\right) \\ &\leq C\left(\sum_{i=1}^{K}\left\|\widehat{a^i}_\varepsilon - \widehat{a^i}\right\|_2^2 + K_N L_N\left\|\widehat{\alpha}_\varepsilon - \widehat{\alpha}\right\|_2^2\right) \\ &\leq CK_N\log^3(N)\varepsilon^2.\end{aligned}$$

   So we obtain

$$\mathcal{D}_{\widehat{h}} - \mathcal{D}_{\widehat{h}_\varepsilon} \leq CK_N\log^3(N)\varepsilon^2. \tag{4.15}$$

2. Upper bound of $2\mathbb{E}\left(\widehat{\mathcal{D}}_{\widehat{h}_\varepsilon} - \widehat{\mathcal{D}}_{\widehat{h}}\right)$

We have $\widehat{\mathcal{D}}_{\widehat{h}_\varepsilon} - \widehat{\mathcal{D}}_{\widehat{h}} = \widehat{\mathfrak{R}}(\widehat{h}_\varepsilon) - \widehat{\mathfrak{R}}(\widehat{h})$, and

$$\widehat{\mathfrak{R}}(\widehat{h}_\varepsilon) - \widehat{\mathfrak{R}}(\widehat{h}) = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\left\{\phi(Z_i^j\widehat{h}_\varepsilon^i(\bar{X}^j)) - \phi(Z_i^j\widehat{h}^i(\bar{X}^j))\right\}. \tag{4.16}$$

For all $j \in [\![1, N]\!]$ and $i \in \mathcal{Y}$, one has $Z_i^j \in \{-1, 1\}$ and,

$$\begin{aligned}\left|\phi(Z_i^j\widehat{h}_\varepsilon^i(\bar{X}^j)) - \phi(Z_i^j\widehat{h}^i(\bar{X}^j))\right| &= \left|\widehat{h}^i - \widehat{h}_\varepsilon^i\right|(\bar{X}^j) \times \left|2 - Z_i^j(\widehat{h}^i(\bar{X}^{(j)}) + \widehat{h}_\varepsilon^i(\bar{X}^j))\right| \\ &\leq 4\left|\widehat{h}^i - \widehat{h}_\varepsilon^i\right|(\bar{X}^j).\end{aligned}$$

Then, we obtain:

$$\begin{aligned}\left|\widehat{\mathfrak{R}}(\widehat{h}_\varepsilon) - \widehat{\mathfrak{R}}(\widehat{h})\right| &\leq \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\left|\phi(Z_i^j\widehat{h}_\varepsilon^i(\bar{X}^j)) - \phi(Z_i^j\widehat{h}^i(\bar{X}^j))\right| \\ &\leq \frac{8}{N}\sum_{j=1}^{N}\sum_{i=1}^{K}\left|\widehat{\pi}_i(\bar{X}^j) - \widehat{\pi}_i^\varepsilon(\bar{X}^j)\right|\end{aligned}$$

and,

$$\widehat{\mathfrak{R}}(\widehat{h}_\varepsilon) - \widehat{\mathfrak{R}}(\widehat{h}) \leq \frac{8K}{N}\sum_{j=1}^{N}\sqrt{\sum_{i=1}^{K}\left|\widehat{F}_i(\bar{X}^j) - \widehat{F}_i^\varepsilon(\bar{X}^j)\right|^2} \tag{4.17}$$

For all $j \in [\![1, N]\!]$ and $i \in [\![1, K]\!]$, one has:

$$\widehat{F}_i(\bar{X}^j) := \sum_{k=0}^{n-1}\left(\frac{\widehat{b}_i}{\widehat{\sigma}^2}(X_{k\Delta}^j)(X_{(k+1)\Delta}^j - X_{k\Delta}^j) - \frac{\Delta}{2}\frac{\widehat{b}_i^2}{\widehat{\sigma}^2}(X_{k\Delta}^j)\right)$$

and,

$$\begin{aligned}\left|\widehat{F}_i(\bar{X}^j) - \widehat{F}_i^\varepsilon(\bar{X}^j)\right|^2 &\leq 2\left(\sum_{k=0}^{n-1}\left(\frac{\widehat{b}_i}{\widehat{\sigma}^2} - \frac{\widehat{b}_{i,\varepsilon}}{\widehat{\sigma}_\varepsilon^2}\right)(X_{k\Delta}^j)(X_{(k+1)\Delta}^j - X_{k\Delta}^j)\right)^2 \\ &\quad + \frac{\Delta^2}{2}\left(\sum_{k=0}^{n-1}\left(\frac{\widehat{b}_i^2}{\widehat{\sigma}^2} - \frac{\widehat{b}_{i,\varepsilon}^2}{\widehat{\sigma}_\varepsilon^2}\right)(X_{k\Delta}^j)\right)^2 \\ &\leq 2\left(\sum_{k=0}^{n-1}\left|\frac{\widehat{b}_i}{\widehat{\sigma}^2} - \frac{\widehat{b}_{i,\varepsilon}}{\widehat{\sigma}_\varepsilon^2}\right|(X_{k\Delta}^j)\left|X_{(k+1)\Delta}^j - X_{k\Delta}^j\right|\right)^2 \\ &\quad + \frac{\Delta}{2}\sum_{k=0}^{n-1}\left|\frac{\widehat{b}_i^2}{\widehat{\sigma}^2} - \frac{\widehat{b}_{i,\varepsilon}^2}{\widehat{\sigma}_\varepsilon^2}\right|^2(X_{k\Delta}^j).\end{aligned}$$

For all $i \in \mathcal{Y}$, we have:

$$\begin{aligned}\left|\frac{\widehat{b}_i}{\widehat{\sigma}^2} - \frac{\widehat{b}_{i,\varepsilon}}{\widehat{\sigma}_\varepsilon^2}\right| = \frac{\left|\widehat{b}_i\widehat{\sigma}_\varepsilon^2 - \widehat{b}_{i,\varepsilon}\widehat{\sigma}^2\right|}{\widehat{\sigma}^2\widehat{\sigma}_\varepsilon^2} &\leq \frac{\left|\widehat{b}_i - \widehat{b}_{i,\varepsilon}\right|}{\widehat{\sigma}^2} + \frac{\left|\widehat{b}_{i,\varepsilon}\right| \times \left|\widehat{\sigma}^2 - \widehat{\sigma}_\varepsilon^2\right|}{\widehat{\sigma}^2\widehat{\sigma}_\varepsilon^2} \\ &\leq \log^4(N)\left|\widehat{b}_i - \widehat{b}_{i,\varepsilon}\right| + \log^8(N)\sqrt{(K_N + M)\log^3(N)}\left|\widehat{\sigma}^2 - \widehat{\sigma}_\varepsilon^2\right| \\ &\leq \log^4(N)\left\|\widehat{b}_i - \widehat{b}_{i,\varepsilon}\right\|_\infty + \log^8(N)\sqrt{(K_N + M)\log^3(N)}\left\|\widehat{\sigma}^2 - \widehat{\sigma}_\varepsilon^2\right\|_\infty \\ &\leq C\log^8(N)\sqrt{K_N\log^3(N)}\varepsilon\end{aligned}$$

where $C > 0$ is a constant, and

$$\begin{aligned}\left|\frac{\widehat{b}_i^2}{\widehat{\sigma}^2} - \frac{\widehat{b}_{i,\varepsilon}^2}{\widehat{\sigma}_\varepsilon^2}\right| &\leq 2\log^4(N)\sqrt{(K_N+M)\log^3(N)}\left|\widehat{b}_i - \widehat{b}_{i,\varepsilon}\right| + \log^8(N)(K_N+M)\log^3(N)\left|\widehat{\sigma}^2 - \widehat{\sigma}_\varepsilon^2\right| \\ &\leq 2\log^4(N)\sqrt{(K_N+M)L\log^3}\left\|\widehat{b}_i - \widehat{b}_{i,\varepsilon}\right\|_\infty + \log_N^8(K_N+M)\log^3\left\|\widehat{\sigma}^2 - \widehat{\sigma}_\varepsilon^2\right\|_\infty \\ &\leq C\log^{11}(N)K_N\varepsilon.\end{aligned}$$

It results that for all $i \in \mathcal{Y}$, we have:

$$\begin{aligned}\left|\widehat{F}_i(\bar{X}^j) - \widehat{F}_i^\varepsilon(\bar{X}^j)\right|^2 \leq\ & \log^{19}(N)(K_N+M)\varepsilon^2\left(\sum_{k=0}^{n-1}\left|X^j_{(k+1)\Delta} - X^j_{k\Delta}\right|\right)^2 \\ &+ \log^{22}(N)(K_N+M)^2\varepsilon^2.\end{aligned}$$

We deduce that for all $j \in [\![1, N]\!]$,

$$\sqrt{\sum_{i=1}^{K}\left|\widehat{F}_i(\bar{X}^j) - \widehat{F}_i^\varepsilon(\bar{X}^j)\right|^2} \leq C\log^8(N)\sqrt{K_N\log^3(N)}\varepsilon\left(\sum_{k=0}^{n-1}\left|X^{(j)}_{(k+1)\Delta} - X^{(j)}_{k\Delta}\right| + \sqrt{K_N\log^3(N)}\right).$$

and from $E.1$ and (4.17), we have:

$$\widehat{\mathfrak{R}}(\widehat{h}_\varepsilon) - \widehat{\mathfrak{R}}(\widehat{h}) \leq C\left(\log^8(N)\sqrt{K_N\log^3(N)}\frac{\varepsilon}{N}\sum_{j=1}^{N}\sum_{k=0}^{n-1}\left|X^j_{(k+1)\Delta} - X^j_{k\Delta}\right| + \log^8(N)\sqrt{K_N\log^3(N)}\varepsilon\right). \tag{4.18}$$

For all $k \in [\![0, n-1]\!]$,

$$\begin{aligned}\left|X^1_{(k+1)\Delta} - X^1_{k\Delta}\right| &\leq \left|\int_{k\Delta}^{(k+1)\Delta} b_Y^*(X_s)ds\right| + \left|\int_{k\Delta}^{(k+1)\Delta}\sigma^*(X_s)dW_s\right| \\ &\leq C\sqrt{K_N\log^3(N)}\Delta + \left|\int_{k\Delta}^{(k+1)\Delta}\sigma^*(X_s)dW_s\right| \\ \sum_{k=0}^{n-1}\left|X^{(1)}_{(k+1)\Delta} - X^{(1)}_{k\Delta}\right| &\leq C\left(\sqrt{K_N\log^3(N)} + \sum_{k=0}^{n-1}\left|\int_{k\Delta}^{(k+1)\Delta}\sigma^*(X_s)dW_s\right|\right).\end{aligned}$$

For N large enough, we obtain,

$$\mathbb{E}\left[\sum_{k=0}^{n-1}\left|X^{(1)}_{(k+1)\Delta} - X^{(1)}_{k\Delta}\right|\right] \leq C\left(\sqrt{K_N\log^3(N)} + \sqrt{\sum_{k=0}^{n-1}\mathbb{E}\left(\int_{k\Delta}^{(k+1)\Delta}\sigma^{*2}(X_s)ds\right)}\right),$$

then

$$\mathbb{E}\left[\sum_{k=0}^{n-1}\left|X^{(1)}_{(k+1)\Delta} - X^{(1)}_{k\Delta}\right|\right] \leq C\sqrt{K_N\log^3(N)}. \tag{4.19}$$

We deduce from Equations (4.19) and (4.18) that there exists a constant $C > 0$ such that

$$2\mathbb{E}\left(\widehat{\mathcal{D}}_{\widehat{h}_\varepsilon} - \widehat{\mathcal{D}}_{\widehat{h}}\right) = 2\mathbb{E}\left[\widehat{\mathfrak{R}}(\widehat{h}_\varepsilon) - \widehat{\mathfrak{R}}(\widehat{h})\right] \leq C\log^{11}(N)K_N\varepsilon \tag{4.20}$$

For any function $h \in \mathfrak{H}$ such that $h^i = 2\bar{\pi} - 1$, we denote by $h = h_{\mathbf{a},\alpha,\mathfrak{p}}$ where $\mathbf{a} = \left(a^{(1)}, \ldots, a^{(K)}\right)$ and $a^{(i)}$, $i \in \mathcal{Y}$ and α are coordinate vectors of functions of the approximations spaces $\mathcal{S}_{K_N}$ and $\widetilde{\mathcal{S}}_{K_N}$. We conclude from Equations (4.14), (4.15) and (4.20) that

$$\begin{aligned}\mathbb{E}\left[\mathcal{D}_{\widehat{h}} - 2\widehat{\mathcal{D}}_{\widehat{h}}\right] &\leq CK_N\log^3(N)\varepsilon^2 + C\log^{11}(N)K_N\varepsilon + \mathbb{E}\left(\mathcal{D}_{\widehat{h}_\varepsilon} - 2\widehat{\mathcal{D}}_{\widehat{h}_\varepsilon}\right) \\ &\leq CK_N\log^3(N)\varepsilon^2 + C\log^{11}(N)K_N\varepsilon + \mathbb{E}\left\{\sup_{(\mathbf{a},\alpha)\in\bar{B}'^{K+1}_\varepsilon,\mathfrak{p}\in\overline{B}_\varepsilon}\mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} - 2\widehat{\mathcal{D}}_{h_{\mathbf{a},\alpha,\mathfrak{p}}}\right\}.\end{aligned}$$

Thus, we have:

$$\mathbb{E}\left[\mathcal{D}_{\widehat{h}} - 2\widehat{\mathcal{D}}_{\widehat{h}}\right] \leq CK_N \log^3(N)\varepsilon^2 + \log^{11}(N)K_N\varepsilon + T_1 \tag{4.21}$$

where $T_1 = \mathbb{E}\left\{\sup_{(\mathbf{a},\alpha)\in \bar{B'}_\varepsilon^{K+1}, \mathfrak{p}\in\bar{B}_\varepsilon} \mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} - 2\widehat{\mathcal{D}}_{h_{\mathbf{a},\alpha,\mathfrak{p}}}\right\}$. For all $u > 0$,

$$\begin{aligned} T_1 &= \int_0^{+\infty} \mathbb{P}\left(\sup_{(\mathbf{a},\alpha)\in \bar{B'}_\varepsilon^{K+1}, \mathfrak{p}\in\bar{B}_\varepsilon} \mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} - 2\widehat{\mathcal{D}}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} \geq t\right) dt \\ &\leq u + \int_u^{+\infty} \mathbb{P}\left(\sup_{(\mathbf{a},\alpha)\in \bar{B'}_\varepsilon^{K+1}, \mathfrak{p}\in\bar{B}_\varepsilon} \mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} - 2\widehat{\mathcal{D}}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} \geq t\right) dt \\ &= u + \int_u^{+\infty} L_t dt \end{aligned}$$

where $L_t = \mathbb{P}\left(\sup_{(\mathbf{a},\alpha)\in \bar{B'}_\varepsilon^{K+1}, \mathfrak{p}\in\bar{B}_\varepsilon} \mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} - 2\widehat{\mathcal{D}}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} \geq t\right)$, and for all $t \geq u$,

$$\begin{aligned} L_t &= \mathbb{P}\left(\bigcup_{(\mathbf{a},\alpha)\in \bar{B'}_\varepsilon^{K+1}, \mathfrak{p}\in\bar{B}_\varepsilon} \left\{\mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} - 2\widehat{\mathcal{D}}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} \geq t\right\}\right) \\ &\leq \sum_{a^1\in\bar{B'}_\varepsilon} \cdots \sum_{a^K\in\bar{B'}_\varepsilon} \sum_{\alpha\in\bar{B'}_\varepsilon} \sum_{\mathfrak{p}\in\bar{B}_\varepsilon} \mathbb{P}\left(\mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} - 2\widehat{\mathcal{D}}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} \geq t\right) \end{aligned}$$

For all $(\mathbf{a},\alpha) \in \bar{B'}_\varepsilon^{K+1}$ and for all $\mathfrak{p} \in \bar{B}_\varepsilon$, one has:

$$\begin{aligned} \widehat{\mathcal{D}}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} = \widehat{\mathfrak{R}}\left(h_{\mathbf{a},\alpha,\mathfrak{p}}\right) - \widehat{\mathfrak{R}}(\bar{h}_*) &= \frac{1}{N}\sum_{j=1}^N \sum_{i=1}^K \left\{(1 - Z_i^j h^i_{\mathbf{a},\alpha,\mathfrak{p}}(\bar{X}^j))^2 - (1 - Z_i^j \bar{h}^i_*(\bar{X}^j))^2\right\} \\ &= \frac{1}{N}\sum_{j=1}^N S_j \end{aligned}$$

with, for each $j \in [\![1, N]\!]$,

$$S_j = \sum_{i=1}^K \left\{(1 - Z_i^j h^i_{\mathbf{a},\alpha,\mathfrak{p}}(\bar{X}^j))^2 - (1 - Z_i^j \bar{h}^i_*(\bar{X}^j))^2\right\} = \ell_{h_{\mathbf{a},\alpha,\mathfrak{p}}}(Z, X) - \ell_{\bar{h}_*}(Z, X)$$

where,

$$\ell_{h_{\mathbf{a},\alpha,\mathfrak{p}}}(Z, X) = \sum_{i=1}^K (1 - Z_i^j h^i_{\mathbf{a},\alpha,\mathfrak{p}}(\bar{X}^j))^2.$$

We deduce for all $j \in [\![1, N]\!]$ that

$$\begin{aligned} \mathbb{E}(S_j^2) &\leq \mathbb{E}\left[\left(\ell_{h_{\mathbf{a},\alpha,\mathfrak{p}}}(Z, X) - \ell_{\bar{h}_*}(Z, X)\right)^2\right] \\ &\leq K\sum_{i=1}^K \mathbb{E}\left[\left(h^i_{\mathbf{a},\alpha,\mathfrak{p}} - \bar{h}^i_*\right)^2 (\bar{X}^j)\left(2 - Z_i^j\left(h^i_{\mathbf{a},\alpha,\mathfrak{p}} - \bar{h}^i_*\right)(\bar{X}^j)\right)^2\right] \\ &\leq 16K\sum_{i=1}^K \mathbb{E}\left[\left(h^i_{\mathbf{a},\alpha,p} - \bar{h}^i_*\right)^2 (\bar{X}^j)\right] \end{aligned}$$

since $|Z_i^j| = 1$ and $|h^i_{\mathbf{a},\alpha,p}|,\ |\bar{h}^i_*| \leq 1$. Then, for all $j \in [\![1, N]\!]$,

$$\mathbb{E}(S_j^2) \leq 16K\sum_{i=1}^K \mathbb{E}\left[\left(h^i_{\mathbf{a},\alpha,\mathfrak{p}} - \bar{h}^i_*\right)^2 (\bar{X}^j)\right] = 16K\left[\mathfrak{R}(h_{\mathbf{a},\alpha,\mathfrak{p}}) - \mathfrak{R}(\bar{h}_*)\right]. \tag{4.22}$$

We have

$$\begin{aligned}\mathfrak{R}(h_{\mathbf{a},\alpha,\mathfrak{p}}) - \mathfrak{R}(\bar{h}_*) &= \mathfrak{R}(h_{\mathbf{a},\alpha,\mathfrak{p}}) - \mathfrak{R}(h^0) + \mathfrak{R}(h^0) - \mathfrak{R}(\widetilde{h}) + \mathfrak{R}(\widetilde{h}) - \mathfrak{R}(\bar{h}_*) \\ &\leq \mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} + C\left(\frac{\log^3(N)}{K_N^2} + \frac{n}{\log(N)}\exp\left(-c\log^2(N)\right)\right).\end{aligned}$$

From Equation (4.22), there exists a constant $C_K > 0$ depending on K such that

$$\forall j \in [\![1, N]\!],\ \ \mathbb{E}\left(S_j^2\right) \leq C_K(\mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} + t) = v,\quad \forall\, t \geq \frac{\log^3(N)}{K_N^2} + \frac{n}{\log(N)}\exp\left(-c\log^2(N)\right).$$

Furthermore, for all $k \geq 3$, $\sum_{j=1}^N \left\|S_j^k\right\| \leq \frac{vk!c^{k-2}}{2}$ where $v = CK(\mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} + t)$ and $c = 64K^2$. Thus, we obtain from the Bernstein inequality that $\forall \mathbf{a}, \alpha \in \bar{B'}_\varepsilon,\ \ \forall \mathfrak{p} \in \bar{B}_\varepsilon,\ \ \forall\, t \geq \sqrt{\Delta}$,

$$\begin{aligned}\mathbb{P}\left(\mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} - 2\widehat{\mathcal{D}}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} \geq t\right) &= \mathbb{P}\left(\mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} - \widehat{\mathcal{D}}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} \geq \frac{\mathcal{D}_{h_{a},\alpha,\mathfrak{p}} + t}{2}\right) \\ &\leq \exp\left(-\frac{N\left(\mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} + t\right)^2/8}{CK\left(\mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} + t\right) + c\left(\mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} + t\right)}\right) \\ &= \exp\left(-\frac{N\left(\mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} + t\right)}{8(CK + c)}\right).\end{aligned}$$

Finally, since $\exp\left(-\frac{N\mathcal{D}_{h_{\mathbf{a},\alpha,p}}}{8(CK+c)}\right) \leq 1$, we obtain that $\forall \mathbf{a}, \alpha \in \bar{B'}_\varepsilon,\ \ \forall \mathfrak{p} \in \bar{B}_\varepsilon$,

$$\forall\, t \geq \frac{\log^3(N)}{K_N^2} + \frac{n}{\log(N)}\exp\left(-c\log^2(N)\right),\ \ \mathbb{P}\left(\mathcal{D}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} - 2\widehat{\mathcal{D}}_{h_{\mathbf{a},\alpha,\mathfrak{p}}} \geq t\right) \leq \exp\left(-\frac{Nt}{8(CK + c)}\right).$$

It results that for all $t \geq \frac{\log^3(N)}{K_N^2} + \frac{n}{\log(N)}\exp\left(-c\log^2(N)\right)$,

$$L_t \leq \mathcal{N}_2\left(\varepsilon, \bar{B'}\right)^{K+1}\mathcal{N}_2\left(\varepsilon, \bar{B}\right)\exp\left(-\frac{Nt}{8(CK + c)}\right)$$

and for all $u \geq \frac{\log^3(N)}{K_N^2} + \frac{n}{\log(N)}\exp\left(-c\log^2(N)\right)$,

$$\begin{aligned}T_1 &\leq u + \int_u^{+\infty} L_t dt = u + \mathcal{N}_2\left(\varepsilon, \bar{B'}\right)^{K+1}\mathcal{N}_2\left(\varepsilon, \bar{B}\right)\int_u^{+\infty}\exp\left(-\frac{Nt}{8(CK + c)}\right)dt \\ &\leq u + \frac{8(CK + c)\mathcal{N}_2\left(\varepsilon, \bar{B'}\right)^{K+1}\mathcal{N}_2\left(\varepsilon, \bar{B}\right)}{N}\exp\left(-\frac{Nu}{8(CK + c)}\right)\end{aligned}$$

where $\mathcal{N}_2\left(\varepsilon, \bar{B'}\right)$ and $\mathcal{N}_2\left(\varepsilon, \bar{B}\right)$ are respective covering numbers of $\bar{B'}$ and $\bar{B}$.

We set

$$u = \frac{\log^3(N)}{K_N^2} + \frac{n}{\log(N)}\exp\left(-c\log^2(N)\right) + \frac{88K\left((K+1)\log\left(\mathcal{N}_2(\varepsilon, \bar{B'})\right) + \log\left(\mathcal{N}_2(\varepsilon, \bar{B})\right)\right)}{N}$$

and we deduce that

$$T_1 \leq C\left(\frac{\log^3(N)}{K_N^2} + \frac{n}{\log(N)}\exp\left(-c\log^2(N)\right) + \frac{\log(\mathcal{N}_2(\varepsilon, \bar{B'})) + \log(\mathcal{N}_2(\varepsilon, \bar{B}))}{N}\right).$$

For every real number $R > 0$, the euclidean ball of radius R denoted by $\bar{B}_2(0, R) \subset \mathbb{R}^d$ has a covering number $\mathcal{N}_2(\varepsilon, \bar{B}_2(0, R))$ that satisfies:

$$\mathcal{N}_2(\varepsilon, \bar{B}_2(0, R)) \leq \left(\frac{3R}{\varepsilon}\right)^d$$

(see [69], Chapter 15 Prop 1.3). We deduce that

$$\mathcal{N}_2(\varepsilon, \bar{B}') \leq \left(\frac{3\sqrt{(K_N+M)\log^3(N)}}{\varepsilon} \right)^{K_N+M}, \quad \mathcal{N}_2(\varepsilon, \bar{B}) \leq \left(\frac{3}{\varepsilon}\right)^K$$

We set $\varepsilon = 1/N\sqrt{(K_N+M)\log^3(N)}$ and we obtain,

$$T_1 \leq C\left(\frac{\log^3(N)}{K_N^2} + \frac{n}{\log(N)} \exp\left(-c\log^2(N)\right) + \frac{K_N \log(N)}{N} \right)$$

and it follows from (4.21) that,

$$\mathbb{E}\left[\mathcal{D}_{\widehat{h}} - 2\widehat{\mathcal{D}}_{\widehat{h}}\right] \leq C\left(\frac{\log^3(N)}{K_N^2} + \frac{n}{\log(N)} \exp\left(-c\log^2(N)\right) + \frac{K_N \log(N)}{N} \right). \tag{4.23}$$

Finally, from Equations (4.23), (4.13) and (4.7), there exists a constant $C > 0$ such that

$$\mathbb{E}\left[\mathfrak{R}(\widehat{h})\right] - \mathfrak{R}(h^*) \leq C\left(\frac{\log^3(N)}{K_N^2} + \frac{K_N \log(N)}{N} + \frac{n}{\log(N)} \exp\left(-c\log^2(N)\right) \right),$$

and the excess risk of the empirical classifier $\widehat{g} = g_{\widehat{h}}$ satisfies:

$$\mathbb{E}\left[\mathcal{R}(\widehat{g})\right] - \mathcal{R}(g^*) \leq C\left(\frac{\log^{3/2}(N)}{K_N} + \left(\frac{K_N \log(N)}{N}\right)^{1/2} + \Delta^{1/2} + \sqrt{\frac{n}{\log(N)}} \exp\left(-\frac{c}{2}\log^2(N)\right) \right).$$

□

4.8.2 Proof of Proposition 4.5.2

Proof. Define for all function $h \in \mathbb{L}^2(\mathbb{R})$ and from the diffusion process $X = (X_t)_{t\in[0,1]}$ the following pseudo-norm:

$$\|h\|_X^2 := \int_0^1 h^2(X_s)ds.$$

Then, from Assumption 4.5.1, we obtain that

$$\underline{\Delta}_b \leq \|b_1^* - b_0^*\|_X \leq \overline{\Delta}_b$$

where $\underline{\Delta}_b$ and $\overline{\Delta}_b$ are positive real number such that $\overline{\Delta}_b > \underline{\Delta}_b$. Let $m \geq 1$ be an integer. Consider the subdivision

$$\Gamma_m := \left\{ t_k = \underline{\Delta}_b + k\frac{\overline{\Delta}_b - \underline{\Delta}_b}{m}, \;\; k = 0, \ldots, m \right\}$$

of the compact interval $[\underline{\Delta}_b, \overline{\Delta}_b]$. Set $f^* = b_1^* - b_0^*$, then $\Delta_{b^*}^2 = \mathbb{E}\left[\int_0^1 f^{*2}(X_s)ds\right]$. Then, from the total probability formula, we have

$$\mathbb{P}\left(0 < \left|\eta(X) - \frac{1}{2}\right| \leq \varepsilon\right) = \sum_{k=0}^{m-1} \mathbb{P}\left(0 < \left|\eta(X) - \frac{1}{2}\right| \leq \varepsilon \mid \|f^*\|_X \in [t_k, t_{k+1}]\right) \mathbb{P}\left(\|f^*\|_X \in [t_k, t_{k+1}]\right)$$

For all $k \in [\![0, m-1]\!]$,

$$\begin{aligned}\mathbb{P}\left(0 < \left|\eta(X) - \frac{1}{2}\right| \leq \varepsilon \mid \|f^*\|_X \in [t_k, t_{k+1}]\right) &= \mathbb{P}\left(0 < \frac{|q_1(X) - q_0(X)|}{2(q_0(X)+q_1(X))} \leq \varepsilon \mid \|f^*\|_X \in [t_k, t_{k+1}]\right) \\ &= T_1^k + T_2^k\end{aligned}$$

where,

$$\begin{aligned}T_1^k &= \mathbb{P}\left(\{q_1(X) < q_0(X)\} \cap \left\{ \frac{|q_1(X) - q_0(X)|}{2(q_0(X)+q_1(X))} \leq \varepsilon \right\} \mid \|f^*\|_X \in [t_k, t_{k+1}] \right) \\ T_2^k &= \mathbb{P}\left(\{q_1(X) > q_0(X)\} \cap \left\{ \frac{|q_1(X) - q_0(X)|}{2(q_0(X)+q_1(X))} \leq \varepsilon \right\} \mid \|f^*\|_X \in [t_k, t_{k+1}] \right).\end{aligned}$$

Upper-bound of T_1^k. The first term T_1 satisfies

$$\begin{aligned} T_1^k &\leq \mathbb{P}\left(\{q_1(X) < q_0(X)\} \cap \left\{\left|\frac{q_1(X)}{q_0(X)} - 1\right| \leq 4\varepsilon\right\} \mid \|f^*\|_X \in [t_k, t_{k+1}]\right) \\ &\leq \frac{1}{2}T_{1,1}^k + \frac{1}{2}T_{1,2}^k \end{aligned}$$

where,

$$\begin{aligned} T_{1,1}^k &= \mathbb{P}\left(\{q_1(X) < q_0(X)\} \cap \left|\frac{q_1(X)}{q_0(X)} - 1\right| \leq 4\varepsilon \mid \|b_1^* - b_0^*\|_X \in [t_k, t_{k+1}], Y = 0\right), \\ T_{1,2}^k &= \mathbb{P}\left(\{q_1(X) < q_0(X)\} \cap \left|\frac{q_1(X)}{q_0(X)} - 1\right| \leq 4\varepsilon \mid \|b_1^* - b_0^*\|_X \in [t_k, t_{k+1}], Y = 1\right). \end{aligned}$$

On the event $\{Y = 0\}$, we have

$$\begin{aligned} \frac{q_1}{q_0}(X) &= \exp\left(\int_0^1 (b_1^* - b_0^*)(X_s)b_0^*(X_s)ds - \frac{1}{2}\int_0^1 (b_1^* - b_0^*)(b_0^* + b_1^*)(X_s)ds + \int_0^1 (b_1^* - b_0^*)(X_s)dW_s\right) \\ &= \exp\left(-\frac{1}{2}\int_0^1 f^{*2}(X_s)ds + \int_0^1 f^*(X_s)dW_s\right) \\ &= \exp\left(-\frac{1}{2}\|f^*\|_X^2 + \int_0^1 f^*(X_s)dW_s\right). \end{aligned}$$

On the event $\{Y = 1\}$, we have

$$\begin{aligned} \frac{q_1}{q_0}(X) &= \exp\left(\int_0^1 (b_1^* - b_0^*)(X_s)b_1^*(X_s)ds - \frac{1}{2}\int_0^1 (b_1^* - b_0^*)(b_0^* + b_1^*)(X_s)ds + \int_0^1 (b_1^* - b_0^*)(X_s)dW_s\right) \\ &= \exp\left(\int_0^1 \frac{f^{*2}(X_s)}{2}ds + \int_0^1 f^*(X_s)dW_s\right) \\ &= \exp\left(\frac{1}{2}\|f^*\|_X^2 + \int_0^1 f^*(X_s)dW_s\right). \end{aligned}$$

We deduce for all $\varepsilon < 1/8$ and for all $k \in [\![0, m-1]\!]$ that

$$\begin{aligned} T_{1,1}^k &= \mathbb{P}\left(\left|\exp\left(-\frac{1}{2}\|f^*\|_X^2 + \int_0^1 f^*(X_s)dW_s\right) - 1\right| \leq 4\varepsilon \mid \|f^*\|_X \in [t_k, t_{k+1}], Y = 0\right) \\ &\leq \mathbb{P}\left(\log(1-4\varepsilon) \leq \int_0^1 f^*(X_s)dW_s - \frac{\|f^*\|_X^2}{2} \leq \log(1+4\varepsilon) \mid \|f^*\|_X \in [t_k, t_{k+1}], Y = 0\right) \\ &\leq \mathbb{P}\left(-\frac{8\varepsilon}{\Delta_{b^*}} + \frac{\|f^*\|_X^2}{2\Delta_{b^*}} \leq \int_0^1 \frac{f^*(X_s)}{\Delta_{b^*}}dW_s \leq \frac{4\varepsilon}{\Delta_{b^*}} + \frac{\|f^*\|_X^2}{2\Delta_{b^*}} \mid \|f^*\|_X \in [t_k, t_{k+1}], Y = 0\right) \\ &\leq \mathbb{P}\left(-\frac{8\varepsilon}{\Delta_{b^*}} + \frac{t_k^2}{2\Delta_{b^*}} \leq \int_0^1 \frac{f^*(X_s)}{\Delta_{b^*}}dW_s \leq \frac{4\varepsilon}{\Delta_{b^*}} + \frac{t_{k+1}^2}{2\Delta_{b^*}} \mid \|f^*\|_X \in [t_k, t_{k+1}], Y = 0\right) \end{aligned}$$

and,

$$\begin{aligned} T_{1,2}^k &= \mathbb{P}\left(\left|\exp\left(\frac{1}{2}\|f^*\|_X^2 + \int_0^1 f^*(X_s)dW_s\right) - 1\right| \leq 4\varepsilon \mid \|f^*\|_X \in [t_k, t_{k+1}], Y = 1\right) \\ &\leq \mathbb{P}\left(\log(1-4\varepsilon) \leq \int_0^1 f^*(X_s)dW_s + \frac{\|f^*\|_X^2}{2} \leq \log(1+4\varepsilon) \mid \|f^*\|_X \in [t_k, t_{k+1}], Y = 0\right) \\ &\leq \mathbb{P}\left(-\frac{8\varepsilon}{\Delta_{b^*}} - \frac{\|f^*\|_X^2}{2\Delta_{b^*}} \leq \int_0^1 \frac{f^*(X_s)}{\Delta_{b^*}}dW_s \leq \frac{4\varepsilon}{\Delta_{b^*}} - \frac{\|f^*\|_2^2}{2\Delta_{b^*}} \mid \|f^*\|_X \in [t_k, t_{k+1}], Y = 0\right) \\ &\leq \mathbb{P}\left(-\frac{8\varepsilon}{\Delta_{b^*}} - \frac{t_{k+1}^2}{2\Delta_{b^*}} \leq \int_0^1 \frac{f^*(X_s)}{\Delta_{b^*}}dW_s \leq \frac{4\varepsilon}{\Delta_{b^*}} - \frac{t_k^2}{2\Delta_{b^*}} \mid \|f^*\|_X \in [t_k, t_{k+1}], Y = 0\right). \end{aligned}$$

Under Assumption 4.5.1, the random variable $Z = (1/\Delta_{b^*}) \int_0^1 f^*(X_s) dW_s$ admits a density function p_Z that is bounded on the real line $\mathbb{R}$. Thus, there exists a constant $C_Z > 0$ such that

$$\forall\, x \in \mathbb{R},\ \ p_Z(x) \leq C_Z.$$

We finally obtain that for all $k \in [\![0, m-1]\!]$,

$$T_{1,1}^k \leq \int_{-\frac{8\varepsilon}{\Delta_{b^*}} + \frac{t_k^2}{2\Delta_{b^*}}}^{\frac{4\varepsilon}{\Delta_{b^*}} + \frac{t_{k+1}^2}{2\Delta_{b^*}}} p_Z(x)dx \leq C\left(\frac{12}{\Delta_{b^*}}\varepsilon + \frac{1}{2}(t_{k+1}^2 - t_k^2)\right) \leq C\left(\frac{12}{\Delta_{b^*}}\varepsilon + \frac{\overline{\Delta}_b - \underline{\Delta}_b}{2m}\right)$$

$$T_{1,2}^k \leq \int_{-\frac{8\varepsilon}{\Delta_{b^*}} - \frac{t_{k+1}^2}{2\Delta_{b^*}}}^{\frac{4\varepsilon}{\Delta_{b^*}} - \frac{t_k^2}{2\Delta_{b^*}}} p_Z(x)dx \leq C\left(\frac{12}{\Delta_{b^*}}\varepsilon + \frac{1}{2}(t_{k+1}^2 - t_k^2)\right) \leq C\left(\frac{12}{\Delta_{b^*}}\varepsilon + \frac{\overline{\Delta}_b - \underline{\Delta}_b}{2m}\right)$$

where $C > 0$ is a new constant. Then, for all $k \in [\![0, m-1]\!]$,

$$T_1^k \leq C\left(\frac{12}{\Delta_{b^*}}\varepsilon + \frac{\overline{\Delta}_b - \underline{\Delta}_b}{2m}\right).$$

Using similar reasoning, we obtain that for all $k \in [\![0, m-1]\!]$,

$$T_2^k \leq C\left(\frac{12}{\Delta_{b^*}}\varepsilon + \frac{\overline{\Delta}_b - \underline{\Delta}_b}{2m}\right).$$

Thus, we conclude that for all $m \geq 1$,

$$\begin{aligned}
\mathbb{P}\left(0 < \left|\eta(X) - \frac{1}{2}\right| \leq \varepsilon\right) &= \sum_{k=0}^{m-1} \mathbb{P}\left(0 < \left|\eta(X) - \frac{1}{2}\right| \leq \varepsilon \mid \|f^*\|_X \in [t_k, t_{k+1}]\right) \mathbb{P}\left(\|f^*\|_X \in [t_k, t_{k+1}]\right) \\
&\leq \sum_{k=0}^{m-1} \left(T_1^k + T_1^k\right) \mathbb{P}\left(\|f^*\|_X \in [t_k, t_{k+1}]\right) \\
&\leq C\frac{12}{\Delta_{b^*}}\varepsilon + \frac{\overline{\Delta}_b - \underline{\Delta}_b}{2m}.
\end{aligned}$$

Finally, we tend m toward infinity and obtain

$$\mathbb{P}\left(0 < \left|\eta(X) - \frac{1}{2}\right| \leq \varepsilon\right) \leq C\frac{12}{\Delta_{b^*}}\varepsilon.$$

□

Bibliography

[1] J. Albert-Smet, A. Torrente, and J. Romo. Band depth based initialization of k-means for functional data clustering. *Adv. Data Anal. Classif.*, 17(2):463–484, 2023. 2

[2] B. Andrew, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999. 40

[3] S. Arora, L. Babai, J. Stern, and Z. Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences*, 54(2):317–331, 1997. 132

[4] J.-Y. Audibert, A.-B. Tsybakov, et al. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007. 27, 85, 95, 100

[5] A. Baíllo, A. Cuevas, and R. Fraiman. Classification methods for functional data. *The Oxford handbook of functional data analysis*, 2011. 84

[6] F.-M. Bandi and P.-C.-B. Phillips. Fully nonparametric estimation of scalar diffusion models. *Econometrica*, 71(1):241–283, 2003. 37

[7] Y. Baraud, F. Comte, and G. Viennet. Model selection for (auto-) regression with dependent data. *ESAIM: Probability and Statistics*, 5:33–49, 2001. 48, 76

[8] P.-L Bartlett, M.-I Jordan, and J.-D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. 130, 132, 136

[9] B. Biais, T. Björk, J. Cvitanić, N. El Karoui, E. Jouini, J.-C. Rochet, and M.-C. Quenez. Non-linear pricing theory and backward stochastic differential equations. *Financial Mathematics: Lectures given at the 3rd Session of the Centro Internazionale Matematico Estivo (CIME) held in Bressanone, Italy, July 8–13, 1996*, pages 191–246, 1997. 2

[10] G. Biau, F. Bunea, and M.-H. Wegkamp. Functional classification in hilbert spaces. *IEEE Transactions on Information Theory*, 51(6):2163–2172, 2005. 17

[11] B. Cadre. Supervised classification of diffusion paths. *Mathematical Methods of Statistics*, 22(3):213–225, 2013. 19, 20, 85, 130, 132, 138

[12] M.-P. Chandra et al. On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India*, volume 2, pages 49–55, 1936. 17

[13] J. Chang. Functional data classification using the rotation forest method combined with the patch selection. *Comm. Statist. Simulation Comput.*, 52(7):3365–3378, 2023. 2

[14] G. Ciolek, D. Marushkevych, and M. Podolskij. On lasso estimator for the drift function in diffusion models. *arXiv preprint arXiv:2209.05974*, 2022. 30

[15] E. Clement. Estimation of diffusion processes by simulated moment methods. *Scandinavian Journal of Statistics*, 24(3):353–369, 1997. 1, 4, 36

[16] A. Cohen, M. Davenport, and D. Leviatan. On the stability and accuracy of least squares approximations. *Foundations of computational mathematics*, 13(5):819–834, 2013. 125

[17] F. Comte. *Nonparametric Estimation*. Spartacus IDH, 2017. 37

[18] F. Comte and V. Genon-Catalot. Nonparametric drift estimation for iid paths of stochastic differential equations. *The Annals of Statistics*, 48(6):3336–3365, 2020. 36, 47, 70, 85, 86, 122

[19] F. Comte and V. Genon-Catalot. Regression function estimation as a partly inverse problem. *Annals of the Institute of Statistical Mathematics*, 72(4):1023–1054, 2020. 40, 86, 125

[20] F. Comte and V. Genon-Catalot. Drift estimation on non compact support for diffusion models. *Stochastic Processes and their Applications*, 134:174–207, 2021. 2, 4, 5, 8, 25, 40, 47, 85, 86

[21] F. Comte, V. Genon-Catalot, et al. Regression function estimation as a partly inverse problem. *Preprint Hal*, 2018. 82

[22] F. Comte, V. Genon-Catalot, Y. Rozenholc, et al. Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli*, 2007. 4, 9, 10, 36, 85

[23] A. Cuevas, M. Febrero, and R. Fraiman. Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496, 2007. 17

[24] S. Dabo-Niang, F. Ferraty, F. Rossi, and N. Villa. Recent advances in the use of svm for functional data classification. *Functional and operatorial statistics*, pages 273–280, 2008. 84

[25] C. De Boor. *A practical guide to splines*, volume 27. springer-verlag New York, 1978. 85

[26] P.-L. De Micheaux, Pavlo Mozharovskyi, and Myriam Vimond. Depth for curve data and applications. *Journal of the American Statistical Association*, 116(536):1881–1897, 2021. 84

[27] L. Della-Maestra and M. Hoffmann. Nonparametric estimation for interacting particle systems: Mckean–vlasov models. *Probability Theory and Related Fields*, 182(1):551–613, 2022. 85

[28] C. Denis, C. Dion-Blanc, E. Ella-Mintsa, and V.-C. Tran. Nonparametric plug-in classifier for multiclass classification of sde paths. *arXiv preprint arXiv:2212.10259*, 2022. 25, 37, 47, 55, 82, 139, 140, 141

[29] C. Denis, C. Dion-Blanc, and M. Martinez. Consistent procedures for multiclass classification of discrete diffusion paths. *Scandinavian Journal of Statistics*, 47(2):516–554, 2020. 4, 12, 15, 20, 85, 87, 88, 100, 130, 132, 135, 138

[30] C. Denis, C. Dion-Blanc, and M. Martinez. A ridge estimator of the drift from discrete repeated observations of the solutions of a stochastic differential equation. *Bernoulli*, 2021. 2, 4, 7, 8, 23, 25, 29, 36, 39, 41, 42, 43, 47, 59, 61, 64, 66, 71, 85, 89, 91, 92, 93, 94, 97, 106, 107, 109, 115, 117, 123, 125

[31] R.-A. DeVore and G.-G. Lorentz. *Constructive approximation*, volume 303. Springer Science & Business Media, 1993. 40

[32] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013. 12, 89

[33] D. Domingo, A. d'Onofrio, and F. Flandoli. Properties of bounded stochastic processes employed in biophysics. *Stochastic Analysis and Applications*, 38(2):277–306, 2020. 84

[34] N. El Karoui, S. Peng, and M. C. Quenez. Backward stochastic differential equations in finance. *Mathematical finance*, 7(1):1–71, 1997. 84

[35] E. Ella-Mintsa. Nonparametric estimation of the diffusion coefficient from sde paths. *arXiv preprint arXiv:2307.03960*, 2023. 29

[36] R. Erban and S. J. Chapman. Stochastic modelling of reaction–diffusion processes: algorithms for bimolecular reactions. *Physical biology*, 6(4):046001, 2009. 84

[37] A. Fermanian. *Learning time-dependent data with the signature transform*. PhD thesis, Sorbonne université, 2021. 17

[38] F. Ferraty and P. Vieu. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44(1-2):161–173, 2003. 17

[39] F. Ferraty and P. Vieu. Reference manual for implementing nonparametric functional data analysis (npfda). *Companion Manual of the Book: Nonparametric Functional Data Analysis, Theory and Practice, Springer-Verlag, New York*, 2006. 1

[40] L. Ferré and N. Villa. Multilayer perceptron with functional inputs: an inverse regression approach. *Scandinavian Journal of Statistics*, 33(4):807–823, 2006. 17

[41] D. Florens. Estimation of the diffusion coefficient from crossings. *Statistical Inference for Stochastic Processes*, 1:175–195, 1998. 4, 36, 37

[42] D. Florens-Zmirou. On estimating the diffusion coefficient from discrete observations. *Journal of applied probability*, 30(4):790–804, 1993. 2, 4, 37, 49, 52, 53

[43] A. Friedman. *Stochastic Differential Equations and Applications*, volume 1. Academic Press, 1975. 131

[44] S. Gadat, S. Gerchinovitz, and C. Marteau. Optimal functional supervised classification with separation condition. *Bernoulli*, 26(3):1797–1831, 2020. 18, 21, 85, 95, 100, 135

[45] S. Gadat, T. Klein, and C. Marteau. Classification in general finite dimensional spaces with the k-nearest neighbor rule. *The Annals of Statistics*, 44(3):982–1009, 2016. 85, 86, 95, 100

[46] V. Genon-Catalot and J. Jacod. On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. In *Annales de l'IHP Probabilités et statistiques*, volume 29, pages 119–151, 1993. 1, 4, 36

[47] V. Genon-Catalot and J. Jacod. Estimation of the diffusion coefficient for diffusion processes: random sampling. *Scandinavian Journal of Statistics*, pages 193–221, 1994. 36

[48] V. Genon-Catalot, T. Jeantheau, and C. Laredo. Parameter estimation for discretely observed stochastic volatility models. *Bernoulli*, pages 855–872, 1999. 36

[49] V. Genon-Catalot, C. Laredo, and D. Picard. Non-parametric estimation of the diffusion coefficient by wavelets methods. *Scandinavian Journal of Statistics*, pages 317–335, 1992. 36

[50] A. K Ghosh and P. Chaudhuri. On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, 32(2):327–350, 2005. 17

[51] A. Gloter. Discrete sampling of an integrated diffusion process and parameter estimation of the diffusion coefficient. *ESAIM: Probability and Statistics*, 4:205–227, 2000. 1, 4, 36

[52] E. Gobet. Lan property for ergodic diffusions with discrete observations. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38:711–737, 2002. 4, 55, 86, 91, 95, 101, 102, 103, 121, 131

[53] C. Gourieroux, H.-T. Nguyen, and S. Sriboonchitta. Nonparametric estimation of a scalar diffusion model from discrete time data: a survey. *Annals of Operations Research*, 256:203–219, 2017. 37

[54] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006. 21, 24, 39, 89, 108, 117, 120

[55] T. Hastie, R. Tibshirani, and J. Friedman. Springer series in statistics the elements of statistical learning data mining, inference. Technical report, and Prediction. Technical report, 2001. 42

[56] M. Hoffmann. Minimax estimation of the diffusion coefficient through irregular samplings. *Statistics & probability letters*, 32:11–24, 1997. 4, 36

[57] M. Hoffmann. Adaptive estimation in diffusion processes. *Stochastic processes and their Applications*, 79(1):135–163, 1999. 36, 37, 85

[58] M. Hoffmann. Lp estimation of the diffusion coefficient. *Bernoulli*, pages 447–481, 1999. 4, 9, 10, 22, 36, 37, 42, 43, 53, 85

[59] S.-M. Iacus. *Simulation and inference for stochastic differential equations: with R examples*. Springer Science & Business Media, 2009. 49, 96

[60] J. Jacod. Random sampling in estimation problems for continuous gaussian processes with independent increments. *Stochastic processes and their applications*, 44(2):181–204, 1993. 1, 4, 36

[61] J. Jacod and A. Shiryaev. *Limit theorems for stochastic processes*, volume 288. Springer Science & Business Media, 2013. 113

[62] G.-M. James and T.-J. Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63:533–550, 2001. 17

[63] I. Karatzas and S. Shreve. *Brownian motion and stochastic calculus*, volume 113. springer, 2014. 4, 15, 86

[64] P. Kidger, J. Foster, X. Li, and T.-J. Lyons. Neural sdes as infinite-dimensional gans. In *International conference on machine learning*, pages 5453–5463. PMLR, 2021. 84

[65] D. Lamberton and B. Lapeyre. *Introduction to stochastic calculus applied to finance*. CRC press, 2011. 1, 2

[66] J.-F. Le Gall. *Mouvement brownien, martingales et calcul stochastique*. Springer, 2013. 2, 113

[67] S.-J. Leon, A. Björck, and W. Gander. Gram-schmidt orthogonalization: 100 years and more. *Numerical Linear Algebra with Applications*, 20(3):492–532, 2013. 124

[68] S. López-Pintado and J. Romo. Depth-based classification for functional data. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 72:103, 2006. 17

[69] G.-G. Lorentz, M.-V. Golitschek, and Y. Makovoz. *Constructive approximation: advanced problems*, volume 304. Springer, 1996. 76, 146

[70] N. Marie and A. Rosier. Nadaraya-watson estimator for iid paths of diffusion processes. *arXiv preprint arXiv:2105.06884*, 2021. 4, 85

[71] P. Massart, E. Nédélec, et al. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006. 130, 132

[72] F. Maturo and R. Verde. Supervised classification of curves via a combined use of functional data analysis and tree-based methods. *Comput. Statist.*, 38(1):419–459, 2023. 2

[73] K. Mosler and P. Mozharovskyi. Fast dd-classification of functional data. *Statistical Papers*, 58(4):1055–1089, 2017. 17

[74] D. Nualart. *The Malliavin calculus and related topics*, volume 1995. Springer, 2006. 136

[75] D.-W. Odell-Scott. Multivariate density estimation: theory, practice, and visualization, 1992. 53

[76] A. Oga and Y. Koike. Drift estimation for a multi-dimensional diffusion process using deep neural networks. *arXiv preprint arXiv:2112.13332*, 2021. 30

[77] J.-Y. Park and B. Wang. Nonparametric estimation of jump diffusion models. *Journal of Econometrics*, 222(1):688–715, 2021. 37

[78] J.-O. Ramsay and B;-W. Silverman. *Applied functional data analysis: methods and case studies*. Springer, 2002. 1

[79] J.-O. Ramsay and B.-W. Silverman. *Fitting differential equations to functional data: Principal differential analysis*. Springer, 2005. 1, 84

[80] R. Renò. Nonparametric estimation of stochastic volatility models. *Economics Letters*, 90(3):390–395, 2006. 37

[81] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*, volume 293. Springer Science & Business Media, 2013. 2, 16, 32, 42, 55, 87, 135

[82] P. Romanczuk, M. Bär, W. Ebeling, B. Lindner, and L. Schimansky-Geier. Active brownian particles: From individual to collective stochastic dynamics. *The European Physical Journal Special Topics*, 202:1–162, 2012. 1

[83] E. Schmisser. Non-parametric estimation of the diffusion coefficient from noisy data. *Statistical inference for stochastic processes*, 15(3):193–223, 2012. 36, 37

[84] E. Schmisser. Penalized nonparametric drift estimation for a multidimensional diffusion process. *Statistics*, 47(1):61–84, 2013. 30

[85] E. Schmisser. Non parametric estimation of the diffusion coefficients of a diffusion with jumps. *Stochastic processes and their applications*, 129(12):5364–5405, 2019. 36, 37

[86] J. Sohn, S. Jeong, Y.-M. Cho, and T. Park. Functional clustering methods for binary longitudinal data with temporal heterogeneity. *Comput. Statist. Data Anal.*, 185:Paper No. 107766, 19, 2023. 2

[87] H. Sørensen. Estimation of diffusion parameters for discretely observed diffusion processes. *Bernoulli*, pages 491–508, 2002. 1, 4, 36

[88] P. Soulier. *Estimation fonctionnelle dans divers cadres de dépendance*. PhD thesis, Paris 11, 1993. 2, 36

[89] A.-B. Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008. 92

[90] J.-W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975. 17

[91] S. Van-de Geer. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *The Annals of Statistics*, pages 1779–1801, 1995. 113

[92] V.-N. Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999. 130, 132

[93] J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. *Annual Review of Statistics and its application*, 3:257–295, 2016. 84

[94] S. Wang, J. Cao, and S.-Y. Philip. Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering*, 34(8):3681–3700, 2020. 84

[95] Y. Yang. Minimax nonparametric classification: Rates of convergence. *IEEE Transactions on Information Theory*, 45(7):2271–2284, 1999. 95

[96] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004. 130, 133