Estimation and testing of mixtures of Hilbert-valued features issued from a continuous dictionary

Clément Hardy

École des Ponts
ParisTech

THÈSE DE DOCTORAT
de l'École des Ponts ParisTech

Estimation and testing of mixtures of Hilbert-valued features issued from a continuous dictionary

École doctorale MSTIC

Spécialité : Mathématiques

Thèse préparée au sein du centre d'enseignement et de recherche en mathématiques et calcul scientifique (CERMICS)

Thèse soutenue le 16 février 2023, par
Clément HARDY

Composition du jury:

Fabrice Gamboa Professeur, Université Paul Sabatier	*Rapporteur*
Clément Marteau Professeur, Université de Lyon	*Rapporteur*
Gabriel Peyré Directeur de recherche, École Normale Supérieure	*Président du jury*
Joseph Salmon Professeur, Université de Montpellier	*Examinateur*
Cristina Butucea Professeur, ENSAE, IP PARIS	*Directrice de thèse*
Jean-François Delmas Professeur, École des Ponts	*Directeur de thèse*
Anne Dutfoy Ingénieur-chercheur, EDF R&D	*Invitée*

Remerciements

Merci à Fabrice Gamboa et à Clément Marteau d'avoir accepté de lire et d'évaluer ce manuscrit. Merci à Gabriel Peyré et Joseph Salmon pour leur participation au jury.

Je ne parviens pas à me rappeler de toutes les raisons qui m'ont poussé à entreprendre une thèse en 2019. S'agissait-il de régler définitivement un vieux compte avec les études, d'enluminer mes paraphes par deux petites lettres élégantes ou de profiter encore trois ans de la cantine des ponts et chaussées ? Tout à la fois ou alors rien de tout cela. A vrai dire, j'étais aussi très enthousiaste à l'idée de pouvoir ajouter une petite poussière à ce que j'avais tenté d'apprendre et de comprendre du collège au lycée, en école d'ingénieur et dans mes cahiers de vacances. Peut-être avais-je bon espoir d'en tirer de la fierté sinon au moins un peu de vanité. Dans tous les cas, les considérations m'ayant amené au doctorat n'avaient rien à voir avec ce qu'il m'a apporté de joies et d'épines. Pour les aspects les plus piquants, il y a eu l'âpreté d'un problème trop dur pour soi, les insomnies tissées de calculs retors, les maux de crâne parfois, la pandémie trop longtemps. Les discussions, la liberté, les tableaux noirs et le contentement lorsqu'une preuve daignait enfin se révéler, sont au contraire ce qu'il recelait de plus doux. S'il avait été un paysage, il eût été vallonné à coup sûr. J'ai vu des joies se dissiper sitôt que leur objet était rédigé au propre et en détail : sous les pétales, l'épine ! Combien de preuves aperçues en fin d'après midi m'ont fait faux bond le matin suivant. S'ensuivait un petit déjeuner difficile où tartines et gribouillages de coin de table s'enchevêtraient. Malgré mes pérégrinations par monts et par vaux, je dirai par-dessus tout que cette thèse fut agréable à écrire. J'en aurai un souvenir heureux tant je fus bien entouré. Jean-François, Anne et Cristina ont rendu accessible un problème qui eût été sans doute hors de ma portée sans leur disponibilité et leur implication. J'ai été auprès d'eux un apprenti trois ans durant. D'abord, ils m'ont offert l'écoute et la patience en s'efforçant toujours de comprendre mes explications aux accents d'élucubrations. Ensuite, ils sont intervenus chaque fois que la mathématique se refusait à moi. Je voudrais donc ici les remercier. Leur bienveillance ne sera sans doute pas perceptible à la lecture du manuscrit mais elle a bel et bien infusé dans chaque page.

Le CERMICS a été pour moi un lieu de vie très agréable. On y trouve une farandole de (post-)doctorants drôles et brillants ainsi qu'Isabelle : rassurante et toujours disponible. Sur mon bout de bureau au fond du deuxième étage, je me suis senti à l'aise.

Ce fut pour moi une grande chance de préparer cette thèse au sein d'EDF. J'y ai rencontré des gens, des métiers et tout un tas de problèmes passionnants. Aussi, je remercie tous les membres des départements Périclès, Prisme et MMC avec qui j'ai été amené à collaborer.

Je n'ai pas toujours été assis derrière un bureau, le virus du covid m'a fait prendre l'air. Ce manuscrit a été rédigé, au gré des confinements et des restrictions, chez les uns et les autres. Quand, sous l'effet d'un tropisme inexpliqué, mon ordinateur et mon bloc-note migraient vers la mer, le travail avait des allures de vacances. Bien sûr, amis et famille ne m'ont pas franchement aidé à mettre sur pieds ce manuscrit. A peine se cachaient-ils pour bailler lorsque j'évoquais son contenu. Pourtant, j'ai peine à croire qu'il eût vu le jour si je n'avais pas eu de quoi respirer de temps à autre. Par respirer, j'inclus toutes les modulations possibles du

souffle : de l'inspiration apaisante au grand éclat de rire. Il y a eu les vacances, les concerts, les sorties ici et là. En somme, il y a eu comme une fête en toile de fond. Aux amis et à tous ceux avec qui j'ai pu trinquer, je dis merci. Et puis Sylvie, Benoît et Etienne, cela va de soi, je ne leur ferai pas l'affront d'expliquer pourquoi.

Enfin, il y a eu tous les paysages de Naivasha à Watamu qui ont été pour moi, jusqu'à la fin, une ligne d'horizon et un but. Merci Julie de m'avoir servi de guide une fois le manuscrit terminé.

Abstract

This thesis is devoted to estimation and testing problems for sparse mixtures of features issued from continuous parametric dictionaries. A wide variety of non-linear regression models are considered in a unified framework. In this thesis, the observations are random elements of an Hilbert space resulting from the sum of a deterministic signal containing information, and a noise. The signal is a linear combination (or mixture) of a finite, but possibly increasing, number of features continuously parameterized by a non-linear parameter. We consider a wide range of continuous dictionaries, observation spaces and additive Gaussian noises (white or colored).

One of the main goals of this thesis is to estimate the linear coefficients as well as the non-linear parameters of the mixture in the presence of noise. In the case where only one signal is observed, we propose estimators that are solutions to an optimization problem. In order to quantify the performances of these estimators with respect to the quality of the observations, we establish prediction and estimation bounds that stand with high probability. We show that when the non-linear parameters are sufficiently separated with respect to a Riemannian metric defined by the dictionary, the signal reconstruction almost reaches (up to a logarithmic factor) the performances obtained by the Lasso estimator in the linear case where the features parameters are known and do not need to be estimated. We give refinements of these results for some dictionaries depending on a scaling parameter. We illustrate our results with the Gaussian spikes deconvolution model and with the recontruction of point sources convolved with a low-pass filter.

In practice, it is common to have a set of observations (possibly a continuum) sharing a common structure. We will assume that the signals share an underlying structure by saying that the union of active features in the data set is finite. The question arises whether the estimation of signals can be improved by taking advantage of their common structure. We show in this thesis that, under separation conditions between the non-linear parameters, this improvement occurs. To do so, we define estimators whose performances reach that of the group-Lasso estimator in the multi-task linear regression model where the non-linear parameters are known and do not need to be estimated.

Next, we test whether a noisy observation is derived from a given signal and give non-asymptotic upper bounds for the associated testing risk. In particular, our test encompasses the signal detection framework. We derive an upper bound for the strength that a signal must have in order to be detected in the presence of noise. It turns out that, in this framework, our upper bound on the strength corresponds (up to a logarithmic factor) to the lower bound on the separation rate for signal detection in the high-dimensional linear model associated to a finite dictionary of features. We also propose a procedure to test whether the features of the observed signal belong to a given finite collection. A non-asymptotic bound on the testing risk is given.

Finally, we propose a new numerical approach, using our estimators, to automatically and simultaneously analyze a set of infrared spectra modeled by linear combinations of peaks whose shape and position are parameterized. We study the numerical performances of the proposed algorithm on infrared spectra of polychloroprene rubbers used in a marine environ-

ment.

Keywords: continuous dictionary, high-dimensional regression, mixture models, multi-task learning, non-linear regression, sparsity, spikes deconvolution.

Résumé

Cette thèse aborde des problèmes d'estimation et de test pour des mélanges parcimonieux de composantes issues de dictionnaires continûment paramétrés. Une grande variété de modèles de régression non-linéaires sont considérés dans un cadre unifié. Dans cette thèse, les observations sont des éléments aléatoires d'un espace de Hilbert résultant de la somme d'un signal déterministe, contenant de l'information, et d'un bruit. Le signal est issu d'une combinaison linéaire (ou mélange) d'un nombre fini, mais éventuellement croissant, de composantes continûment paramétrées par un paramètre non-linéaire. Nous considérons un large panel de dictionnaires continus, d'espaces d'observations et de bruits additifs gaussiens (blanc ou colorés).

L'un des buts principaux de cette thèse est d'estimer en présence de bruit les coefficients linéaires ainsi que les paramètres non-linéaires du mélange. Dans le cas où un seul signal est observé, nous proposons des estimateurs solutions d'un problème d'optimisation. Afin de quantifier les performances de ces estimateurs en fonction de la qualité des observations, nous établissons des bornes de prédiction et d'estimation valables en grande probabilité. Nous montrons que lorsque les paramètres non-linéaires sont suffisamment séparés au sens d'une métrique riemannienne définie par le dictionnaire, la reconstruction du signal atteint quasiment (à un facteur logarithmique près) les performances obtenues par l'estimateur Lasso dans le cas linéaire où les paramètres des composantes sont connus et n'ont pas besoin d'être estimés. Nous donnons des raffinements de ces résultats pour certains dictionnaires dépendant d'un paramètre d'échelle. Nous illustrons nos résultats à l'aide du modèle de déconvolution de pics gaussiens et du modèle de reconstruction de sources ponctuelles filtrées.

En pratique, il est fréquent de disposer d'un ensemble d'observations (éventuellement un continuum) partageant une structure commune. Nous supposerons que les signaux partagent une structure sous-jacente en disant que l'union des composantes actives dans l'ensemble des données est finie. La question se pose de savoir si l'estimation des signaux peut être améliorée en tirant parti de leur structure commune. Nous montrons dans cette thèse que, sous des conditions de séparation entre les paramètres non-linéaires, cette amélioration a lieu. Pour ce faire, nous définissons des estimateurs dont les performances atteignent celles de l'estimateur group-Lasso dans le modèle de régression linéaire multi-tâches où les paramètres non linéaires sont connus et n'ont pas besoin d'être estimés.

Ensuite, nous testons si une observation bruitée dérive d'un signal donné et donnons des bornes supérieures non asymptotiques pour le risque de test associé. En particulier, notre test englobe le cadre de la détection de signaux. Nous déduisons une borne supérieure pour l'intensité minimale qu'un signal doit avoir afin d'être détecté en présence de bruit. Il s'avère que, dans ce cadre, notre borne supérieure sur l'intensité minimale correspond (à un facteur logarithmique) à la borne inférieure de la vitesse de séparation pour la détection de signaux dans le modèle linéaire de grande dimension associé à un dictionnaire fini de composantes. Nous proposons également une procédure permettant de tester si les composantes du signal observé appartiennent à une collection finie donnée. Une borne non asymptotique sur le risque de test est donnée.

Enfin, nous proposons une nouvelle approche numérique, utilisant nos estimateurs, pour

analyser automatiquement et simultanément un ensemble de spectres infrarouges modélisés par des combinaisons linéaires de pics dont la dispersion et la position sont paramétrées. Nous étudions les performances numériques de l'algorithme proposé sur des spectres infrarouges de revêtements en polychloroprène vieillis en milieu marin.

Mots clés: dictionnaire continu, parcimonie, régression en grande dimension, régression non-linéaire, modèles de mélange, apprentissage simultané, déconvolution de pics.

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

Contents

1.1 Modélisation statistique

La modélisation statistique s'avère être un outil puissant dès lors que l'on souhaite extraire de l'information d'observations partielles et bruitées. Dans la plupart des domaines des sciences de l'ingénieur, les praticiens sont amenés à manipuler des jeux de données. Par jeu de données, nous désignons un certain nombre d'éléments appartenant à un espace d'observation. Il peut s'agir, entre autres, de spectres infrarouges, d'images ou de séquences d'ADN. Selon la nature des données, l'information d'intérêt à extraire varie. On peut vouloir détecter une bande d'absorption dans un spectre infrarouge afin de révéler la présence d'un composé chimique dans un matériau ([Aragoni et al., 1995], [Butucea et al., 2021]), restaurer une image endommagée ([Mairal et al., 2008], [Mairal et al., 2009]) ou identifier des motifs dans une séquence d'ADN ([Conlon et al., 2003]). En statistiques, on suppose que les données sont générées par une distribution de probabilité sur l'espace d'observation. La modélisation statistique consiste alors à définir une famille de distributions à laquelle est susceptible d'appartenir celle ayant généré les données. Extraire de l'information des données peut consister à retrouver cette dernière ou à vérifier qu'elle satisfait certaines propriétés. Ceci conduit à deux aspects fondamentaux de l'inférence statistique: l'estimation et le test. Dans cette thèse nous supposerons que les données sont générées par une distribution issue d'une famille paramétrique de grande dimension.

1.1.1 Modèles statistiques de grande dimension

En raison des progrès constants en matière de capacité de stockage et de puissance de calcul, les données manipulées aujourd'hui sont souvent issues de modèles statistiques associés à des espaces d'observation et de paramètres aux dimensions arbitrairement grandes. Certains paradigmes de la statistique en faible dimension se révèlent insuffisants dans ce cadre. Classiquement, pour éclairer la compréhension d'un modèle, des résultats asymptotiques sont établis sous l'hypothèse que la dimension de l'espace des paramètres est fixe et que la dimension de l'espace d'observation augmente avec la quantité de données ou la qualité des observations. Malheureusement, ces résultats asymptotiques se révèlent inutiles lorsque la

dimension de l'espace des paramètres est de l'ordre de celle de l'espace d'observation. Nous tâcherons dans cette thèse de fournir des résultats adaptés à la grande dimension. Évidemment, cette approche ne vient pas sans un lot de difficultés à lever. Considérons l'un des modèles statistiques les plus simples, à savoir le modèle de régression linéaire. Ce modèle est pertinent lorsque la relation entre un élément observé et des variables explicatives est linéaire à un bruit additif près (souvent supposé gaussien). Une question naturelle se pose: dans quelle mesure pouvons-nous estimer la relation linéaire entre l'élément observé et les variables explicatives? En grande dimension et sans hypothèses supplémentaires, il est *a priori* impossible de retrouver les coefficients linéaires à partir de l'observation. En effet, même en l'absence de bruit, les coefficients linéaires sont alors solutions d'un système d'équations sous-déterminé. Pour répondre à ce problème, il est d'usage de supposer que le nombre de variables explicatives ayant un impact non nul sur l'observation est inférieur à la dimension de l'espace d'observation. Le modèle devient un modèle de régression linéaire parcimonieux pour lequel de nombreux estimateurs sont disponibles à condition que les variables explicatives actives ne soient pas trop corrélées entre elles (voir [Bühlmann and van de Geer, 2011]). Ce type de phénomène propre à la grande dimension est présent dans de nombreux modèles tels que les modèles graphiques, l'analyse en composantes principales, la régression matricielle (voir [Wainwright, 2019] dans ce sens). Dans ces conditions, l'estimation et le test nécessitent le développement d'outils statistiques complémentaires aux outils classiques.

1.1.2 Modèles de mélange de composantes

Les observations réalisées en pratique ont souvent une structure sous-jacente : une photographie est la superposition de *patches*, un spectre infrarouge est une combinaison linéaire de bandes d'absorption et une séquence d'ADN contient des motifs. Plus généralement, supposons que nous observions un élément y d'un espace d'observation H composé d'un signal porteur d'information et d'un bruit additif. Une structure sous-jacente parcimonieuse peut être identifiée en le décomposant en une combinaison linéaire (ou mélange) de composantes (*features*). En procédant ainsi, une observation complexe de grande dimension peut-être réduite à quelques coefficients linéaires. Ces trois dernières décennies, la décomposition parcimonieuse de signaux, souvent nommée acquisition comprimée (ou codage parcimonieux, voir [Olshausen and Field, 1997]), a suscité un grand intérêt dans la littérature. Pour une introduction à ce domaine de recherche, on peut se référer au papier [Candès and Wakin, 2008] et à l'ouvrage [Foucart and Rauhut, 2013].

Décomposition dans un dictionnaire

Considérons que l'espace d'observation H est un espace normé complet quelconque doté d'un produit scalaire (ou espace de Hilbert) de sorte que la plupart des calculs standards (dérivation, intégration) soient disponibles et que l'on puisse définir une orthogonalité. Les composantes sous-jacentes au signal peuvent provenir d'un dictionnaire (ou collection) fini $\{\varphi_1, \cdots, \varphi_K\}$ d'éléments de H. Ces composantes peuvent être vues comme des variables explicatives, réduisant ainsi le modèle à une régression linéaire. Celle-ci est ensuite qualifiée de grande dimension lorsque la taille du dictionnaire K est supérieure à la dimension de l'espace H. Dans ce cas, le signal se décompose comme suit:

$$y = \sum_{i \in S} \beta_i \varphi_i + w,$$

où $(\beta_i,\, i \in \{1, \cdots, K\})$ sont des coefficients linéaires, $S = \{i,\, \beta_i \neq 0\}$ est l'ensemble des indices des composantes actives et w est un bruit. L'approche statistique consiste à supposer que le bruit w s'ajoutant au signal est distribué selon une certaine loi de probabilité. En supposant que la loi du bruit est connue, l'observation y est générée par une distribution de probabilité appartenant à un ensemble $(\mathbb{P}_\beta,\, \beta \in \mathbb{R}^K)$ paramétré sur un espace de grande

dimension. En général, deux hypothèses fondamentales sont formulées afin d'estimer les coefficients linéaires du mélange. Premièrement, on suppose que le signal a une structure parcimonieuse: parmi les composantes du dictionnaire, seules quelques-unes sont actives, de sorte que le cardinal de S soit inférieur à la dimension de H. On émet ensuite une hypothèse sur la cohérence du dictionnaire en supposant que les composantes actives dans le mélange ne sont pas trop corrélées deux à deux. L'estimation des coefficients linéaires du mélange est ainsi rendue possible. Dès lors, on peut reconstruire un signal de grande dimension à partir d'un petit nombre de coefficients. L'acquisition comprimée permet de stocker des données numériques de manière plus efficace. En outre, si y est une image bruitée, l'estimation du signal permet de supprimer le bruit et de retrouver l'image originale, voir [Elad and Aharon, 2006] pour une présentation plus complète de cette application.

Dictionnaires prédéfinis

Il est légitime de s'interroger sur le choix du dictionnaire utilisé pour modéliser l'observation. Dans diverses applications, on utilise traditionnellement des dictionnaires prédéfinis. Une idée naturelle est de prendre une base de l'espace d'observation H. On a alors affaire à des dictionnaires complets. Les dictionnaires complets comprennent les bases d'ondelettes, les fonctions de base de Fourier...etc. Nous renvoyons à [Mallat, 2009] pour la présentation d'une grande diversité de dictionnaires. Ces trois dernières décennies, l'idée d'utiliser plus d'éléments dans le dictionnaire que la dimension de H s'est révélée particulièrement féconde (on parle alors de dictionnaire redondant). En procédant ainsi, le dictionnaire ne forme plus une base mais offre la possibilité de représenter le signal de manière plus parcimonieuse (voir [Olshausen and Field, 1997]). De tels dictionnaires peuvent être obtenus en combinant plusieurs dictionnaires complets (voir par exemple [Gribonval and Nielsen, 2003]) ou en utilisant une famille de composantes corrélées entre elles. Bien sûr, en procédant ainsi, la décomposition du signal n'est plus unique. On peut donc se demander quelle est la meilleure représentation d'un signal. Il est clair qu'il n'y a pas de réponse simple à cette question. Cela dépend de la tâche que l'on veut accomplir (compression, dé-bruitage...) et de la nature des observations, voir [Rubinstein et al., 2010].

Appprentissage de dictionnaire

Un certain nombre de travaux de recherche se sont concentrés sur l'apprentissage de dictionnaire à partir des données. Typiquement, si nous disposons d'un ensemble d'apprentissage de n observations $(y^{(i)}, 1 \leq i \leq n)$, on peut effectuer une analyse en composantes principales. Les vecteurs propres trouvés par cette procédure fournissent alors un dictionnaire complet. Des méthodes plus complexes ont été développées, comme la décomposition K-SVD. Certains travaux ont également utilisé des dictionnaires paramétriques, voir [Yaghoobi et al., 2009]. L'apprentissage par dictionnaire a donné naissance à un champ de recherche fructueux, citons les travaux de [Lee et al., 2006] et [Duarte-Carvajalino and Sapiro, 2009] pour l'acquisition comprimée ainsi que ceux de [Elad and Aharon, 2006] et [Mairal et al., 2012] pour la restauration d'images bruitées. Nous renvoyons à [Rubinstein et al., 2010] pour un inventaire sur les méthodes d'apprentissage de dictionnaire.

Dictionnaires continûment paramétrés

Lorsque le signal a une structure connue, l'utilisation de dictionnaires appris ou prédéfinis peut manquer une partie de l'information. Supposons par exemple que nous observions une combinaison linéaire de pic translatés, comme c'est le cas en spectroscopie infrarouge. Plutôt que de décomposer le signal dans un dictionnaire quelconque, il est plus intéressant de considérer un dictionnaire composé de pics continûment paramétrés par leurs positions. On utilise alors un dictionnaire continu $(\varphi(\theta), \theta \in \Theta)$ où Θ est l'espace des positions et $\varphi(\theta)$ est un pic centré en θ. Nous nous référons à [Duval and Peyré, 2015] pour le traitement de ce

modèle. Dans ce cadre, le dictionnaire est composé d'un nombre non dénombrable d'éléments fortement corrélés. Nous remarquons que de nombreux signaux peuvent être décomposés dans un dictionnaire continu. En particulier, une base d'ondelettes peut être considérée comme une sous famille d'un dictionnaire continu où φ est une fonction d'échelle et l'espace Θ est le produit cartésien des ensembles continus des paramètres d'échelle et de translation.

Citons parmi les nombreuses références sur les dictionnaires continus, le travail précurseur de [Mallat and Zhang, 1993] pour la décomposition atomique temps-fréquence, [Ekanadham et al., 2011] pour les signaux invariants en translation, [Duval and Peyré, 2015] pour la déconvolution de pics, [Tang et al., 2013b] et [Candès and Fernandez-Granda, 2014] pour la super-résolution.

Exemples de dictionnaires continus

- **Déconvolution de pics.** Le modèle de déconvolution de pics est motivé par des applications en spectroscopie infrarouge (voir [Butucea et al., 2021]). La spectroscopie infrarouge est utilisée dans l'industrie pour le contrôle non-destructif de matériaux. En mesurant l'interaction de rayonnements infrarouges avec la matière, les composants chimiques présents peuvent être identifiés. Les spectres obtenus en spectroscopie infrarouge par transformée de Fourier (IRTF) sont des combinaisons linéaires de pics d'absorption caractérisés par leurs amplitudes, leurs dispersions et leurs positions (voir figure 1.1). Un composant chimique est identifié par la position et la dispersion d'un pic. Il est d'autant plus concentré dans le matériau que l'amplitude du pic associé est élevée. En modélisant un pic à l'aide d'une fonction continûment dilatée et translatée, on peut décomposer un spectre dans un dictionnaire continu. On peut par exemple modéliser un pic d'absorption par une fonction gaussienne paramétrée par sa moyenne et son écart type:

$$\begin{aligned} \varphi \colon \Theta \subset \mathbb{R}^2 &\to L^2(\mathbb{R}) \\ (\mu, \nu) &\mapsto \mathrm{e}^{-\frac{(\cdot-\mu)^2}{2\nu^2}}, \end{aligned}$$

où $L^2(\mathbb{R})$ désigne l'ensemble des fonctions de carré intégrable sur $\mathbb{R}$ relativement à la mesure de Lebesgue.

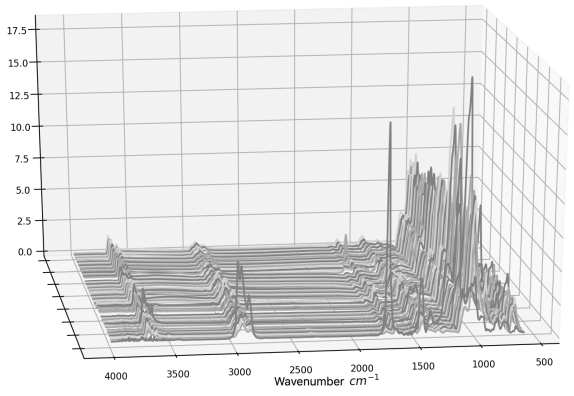

Figure 1.1 – Spectres infrarouges de revêtements en polychloroprène.

Étudier la composition chimique d'un matériau consiste ensuite à estimer les positions et les amplitudes des pics à partir du spectre et à tester si les positions des pics appartiennent à une liste donnée (voir table 1.1).

Table 1.1 – Correspondances entre positions de pics d'absorption et composés chimiques dans un échantillon de polychloroprène. Table issue de [Tchalla, 2017].

Wavenumbers (cm^{-1})	Peak assignment
3690-3400-3364 3200-3014	-OH
2952-2920-2850	$\nu - CH_2, CH_3$ Aliphatic
1731	$\nu - C = O$
1647	$\nu - C = C$ of $HC = CH_2$
1540	$\nu - C = C$ of $R - CR = CH - R$ and $\delta - CH_2$ Aliphatic
1419	$\delta - CH_2, \delta - CH$ Aliphatic
1160-1082	$\nu - Si - O$ (SiO_2)
1009-909	$\nu - Si - O$ $(Si - OH)$
825	$C - Cl$
664	CH Aromatic

- **Super-résolution.** La super-résolution vise à retrouver un signal de résolution fine à partir d'une observation (potentiellement bruitée) de résolution plus grossière issue d'un système d'acquisition. Ce domaine trouve, entre autres, des applications en astronomie et en spectroscopie ([Puschmann and Kneer, 2005], [Harris et al., 1994]). Un exemple de problème de super-résolution considéré dans [Duval and Peyré, 2017a] et [Candès and Fernandez-Granda, 2013] consiste à retrouver des sources ponctuelles, modélisées par une somme pondérée de mesures de Dirac sur le tore $\mathbb{R}/\mathbb{Z}$, à partir de leur convolution avec un filtre passe-bas, en l'occurrence un noyau de Dirichlet de fréquence de coupure $f_c \in \mathbb{N}^*$. Dans cet exemple, l'observation se décompose dans un dictionnaire continu généré par:

$$\begin{aligned} \varphi \colon \mathbb{R}/\mathbb{Z} &\to L^2(\mathbb{R}/\mathbb{Z}) \\ \theta &\mapsto \sum_{k=-f_c}^{f_c} \mathrm{e}^{2i\pi k(\theta - \cdot)} = \frac{\sin\Big((2f_c+1)\pi(\theta-\cdot)\Big)}{\sin\Big(\pi(\theta-\cdot)\Big)}, \end{aligned}$$

  où $L^2(\mathbb{R}/\mathbb{Z})$ désigne les fonctions de carré intégrable sur le tore relativement à la mesure de Lebesgue.

Mélanges issus d'un dictionnaire continu

Nous présentons dans cette section le modèle statistique étudié dans la thèse. Nous supposons que l'observation y appartient à un certain espace de Hilbert H et résulte de la somme d'un signal et d'un bruit. Ce peut être un processus aléatoire continu ou un processus à temps discret. La partie signal est une combinaison linéaire d'un nombre inconnu s de composantes régulières et continûment paramétrées ($\varphi(\theta) \in H, \theta \in \Theta$) où Θ est l'espace des paramètres. Le bruit w est une variable aléatoire à valeurs dans H. Il caractérise avec l'espace d'observation H la qualité de l'information. Dans cette thèse, nous étudierons des suites de modèles dans lesquelles la qualité de l'information croît. Pour ce faire, nous introduisons un paramètre T croissant avec celle-ci et indexons l'espace d'observation ainsi que le bruit par T. On note désormais H_T l'espace d'observation et w_T le bruit. Par exemple, pour un espace d'observation $H_T = \mathbb{R}^T$, T correspond à la dimension. Pour plus de généralité, nous permettons également au dictionnaire de dépendre de T. Le modèle considéré s'écrit pour

des coefficients linéaires $\beta^\star = (\beta_1^\star, \cdots, \beta_s^\star) \in (\mathbb{R}^*)^s$ et des paramètres $\vartheta^\star = (\theta_1^\star, \cdots, \theta_s^\star) \in \Theta^s$:

$$y = \sum_{k=1}^{s} \beta_k^\star \, \phi_T(\theta_k^\star) + w_T \quad \text{dans} \quad H_T, \tag{1.1}$$

où les composantes sont normalisées, à savoir pour tout $\theta \in \Theta$,

$$\phi_T(\theta) = \varphi_T(\theta) / \|\varphi_T(\theta)\|_T, \tag{1.2}$$

avec $\|\cdot\|_T$ la norme dérivée du produit scalaire de l'espace de Hilbert H_T et où par convention, lorsque la parcimonie du signal s est nulle, on a $\beta^\star = 0$, $\vartheta^\star = 0$ et $y = w_T$. On supposera que le bruit w_T est une variable aléatoire dans H_T et qu'il existe une constante $\sigma > 0$ ainsi qu'une quantité $\Delta_T > 0$ telles que:

Hypothèse 1.1.1 (Admissibilité du bruit)**.** *Pour tout $f \in H_T$, la variable aléatoire $\langle f, w_T \rangle_T$ est une variable aléatoire gaussienne centrée satisfaisant:*

$$\mathrm{Var}\left(\langle f, w_T \rangle_T\right) \leq \sigma^2 \, \Delta_T \, \|f\|_T^2.$$

Une grande variété de bruits gaussiens, blanc et colorés, satisfont cette hypothèse, voir les exemples de la section 2.1.2.

Exemple 1.2 (Processus à temps discret sur une grille régulière). Considérons un processus à valeurs réelles y observé sur une grille régulière $t_1 < \cdots < t_T$ d'un intervalle symétrique $[a_T, b_T]$, avec $T \geq 2$, $a_T = -b_T < 0$, $t_j = a_T + j\Delta_T$ pour $j = 1, \ldots, T$ et un pas de discrétisation:

$$\Delta_T = \frac{b_T - a_T}{T}.$$

Si toutes les observations ont le même poids, le processus y est vu comme un élément de l'espace de Hilbert $H_T = L^2(\lambda_T)$ des fonctions à valeurs réelles définies sur $\mathbb{R}$ et de carré intégrable par rapport à la mesure λ_T donnée par:

$$\lambda_T(\mathrm{d}t) = \Delta_T \sum_{j=1}^{T} \delta_{t_j}(\mathrm{d}t),$$

où δ_x désigne la masse de Dirac en x. Supposons que bruit est quant à lui donné par des variables aléatoires gaussiennes centrées et corrélées $G_1, \ldots, G_T$ telles que pour $\sigma_1 > 0$:

$$\mathbb{E}[G_j^2] = \sigma_1^2 \quad \text{et} \quad |\mathbb{E}[G_j G_i]| \leq \sigma_1^2 / T \quad \text{pour } j \neq i \text{ dans } \{1,\ldots,\mathrm{T}\}.$$

Dans ce cas, le processus de bruit $w_T(t) = \sum_{j=1}^{T} G_j \mathbf{1}_{\{t_j\}}(t)$, où $\mathbf{1}_A$ représente la fonction indicatrice d'un ensemble A quelconque, vérifie l'hypothèse 1.1.1 avec $\sigma^2 = 2\sigma_1^2$.

Exemple 1.3 (Processus à temps continu sur le tore). Supposons que nous observions un processus y sur le tore $\mathbb{R}/\mathbb{Z}$ muni de la mesure de Lebesgue notée Leb. Dans ce cadre, l'observation y est vue comme un élément de l'ensemble $H = L^2(\mathrm{Leb})$ des fonctions de carré intégrable sur le tore. Supposons que le bruit soit $w_T = \sum_{k \in \mathbb{N}} \sqrt{\xi_k} \, G_k \, \psi_k$, où $(G_k, k \in \mathbb{N})$ sont des variables aléatoires gaussiennes centrées, indépendantes et de variance σ^2, que $(\psi_k, k \in \mathbb{N})$ soit une base orthonormée de $L^2(\mathrm{Leb})$, et que $\xi = (\xi_k, k \in \mathbb{N})$ soit une suite de nombres réels positifs de carré sommable. L'hypothèse 1.1.1 est alors vérifiée et on a:

$$\mathrm{Var}\left(\langle f, w_T \rangle_{L^2(\lambda_T)}\right) = \sigma^2 \sum_{k \in \mathbb{N}} \xi_k \, \langle f, \psi_k \rangle_{L^2(\mathrm{Leb})}^2 \leq \sigma^2 \, \Delta_T \, \|f\|_{L^2(\mathrm{Leb})}^2 \quad \text{avec} \quad \Delta_T = \sup_{k \in \mathbb{N}} \xi_k.$$

Dans cet exemple, le bruit w_T ne dépend du paramètre T que si ξ, et donc Δ_T, dépendent de T. Nous pouvons envisager différents choix pour ξ. Par exemple, notre cadre englobe le bruit blanc tronqué en prenant pour tout $k \in \mathbb{N}$, $\xi_k = T^{-1} \mathbf{1}_{\{1 \leq k \leq T\}}$. On a alors $\Delta_T = 1/T$.

L'objectif de cette thèse est d'estimer à partir de l'observation y les coefficients linéaires $\beta^\star$, les paramètres non linéaires $\vartheta^\star$, ainsi que le signal $\sum_{k=1}^{s} \beta_k^\star \, \phi_T(\theta_k^\star)$. Nous établirons nos résultats théoriques dans le cas où l'espace des paramètres Θ est unidimensionnel. En revanche nous étudierons numériquement l'estimation des paramètres du modèle (1.1) dans le cas où l'espace Θ est multidimensionnel.

En pratique, il est courant de disposer d'un ensemble d'observations. On peut considérer l'exemple d'un système d'acquisition effectuant une série de mesures dans H_T selon un paramètre de temps ou d'espace. Dans ce cas, la base de données Y se compose d'un ensemble d'observations $Y(z)$ indexées par l'ensemble $\mathcal{Z}$. Ce dernier est alors discret ou continu selon le mode d'acquisition. En pratique, il peut être intéressant d'associer à chaque observation $Y(z)$ un poids renseignant, par exemple, sur la fiabilité de l'observation. Pour ce faire, on peut doter l'ensemble $\mathcal{Z}$ d'une tribu $\mathcal{F}$ et d'une mesure ν de sorte que $(\mathcal{Z}, \mathcal{F}, \nu)$ soit un espace mesuré. En prenant $\mathcal{Z} = \{1, \cdots, n\}$ et ν égale à la mesure de comptage, on peut par exemple considérer une collection de n observations dans H_T. On peut aussi mettre des poids sur les observations en prenant ν égale à la somme pondérée de mesures de Dirac situées sur les éléments de $\mathcal{Z}$. Dans ce cadre, on observe un élément Y dans l'ensemble $L_T = L^2(\nu, H_T)$ des fonctions f, définies sur $(\mathcal{Z}, \mathcal{F}, \nu)$ et à valeurs dans H_T, telles que $\|f\|_{L_T} = \sqrt{\int_{\mathcal{Z}} \|f(z)\|_T^2 \, \nu(\mathrm{d}z)}$ est fini. On généralise alors le modèle (1.1) en posant:

$$Y = \sum_{k=1}^{s} B_k^\star \, \phi_T(\theta_k^\star) + W_T \quad \text{dans} \quad L_T, \tag{1.3}$$

où les coefficients linéaires sont obtenus par l'application $B^\star : z \mapsto B^\star(z) = (B_1^\star(z), \ldots, B_s^\star(z))$ de $L^2(\nu, \mathbb{R}^s)$ et le bruit W_T est une variable aléatoire à valeurs dans L_T.

Les signaux partagent une structure sous-jacente, c'est-à-dire que l'union des composantes actives dans l'ensemble des signaux est finie et égale à s. La question se pose de savoir si l'estimation des signaux peut être accélérée en tirant parti de leur structure commune. Nous montrerons dans cette thèse que sous certaines conditions cette accélération se vérifie.

1.4 Estimation et reconstruction de signaux

Les travaux entrepris dans cette thèse portent sur l'estimation des coefficients linéaires et des paramètres non-linéaires apparaissant dans le modèle (1.1) ainsi que dans sa généralisation (1.3).

Supposons dans un premier temps que l'on dispose d'une observation y issue du modèle (1.1). L'estimation consiste à définir un trio de fonctions mesurables de y, noté $(\hat{s}, \hat{\beta}, \hat{\vartheta}) \in \bigcup_{k \in \mathbb{N}} k \times \mathbb{R}^k \times \Theta^k$ (avec la convention $\mathbb{R}^0 = \{0\}$ et $\Theta^0 = \{0\}$), approximant respectivement le nombre s de composantes actives, les coefficients linéaires $\beta^\star$ et les paramètres non linéaires $\vartheta^\star$ du signal. De manière analogue, on peut définir des estimateurs $(\hat{s}, \hat{B}, \hat{\vartheta}) \in \bigcup_{k \in \mathbb{N}} k \times L^2(\nu, \mathbb{R}^k) \times \Theta^k$ pour le modèle (1.3). Une fois ces estimateurs définis, l'enjeu est de quantifier leur proximité avec les objets qu'ils doivent approcher. Dans ce but, il est nécessaire de se donner des risques d'estimation. Notons que le couple d'estimateur $(\hat{\beta}, \hat{\vartheta})$ est défini à une permutation jointe près sur les composantes de ses deux éléments. En outre, *a priori* le nombre $\hat{s}$ de coefficients de $\hat{\beta}$ est différent de s. Dès lors, une question se pose : le couple estimé $(\hat{\beta}_\ell, \hat{\theta}_\ell)$ avec $\ell \in \{1, \cdots, \hat{s}\}$ approche quel couple $(\beta_k^\star, \theta_k^\star)$ de vrais paramètres ? Pour y répondre, il convient, par exemple, de définir des petits voisinages de taille fixée $r > 0$, relativement à une distance $\mathfrak{d}_T$ dépendant de H_T et φ_T, autour des paramètres non linéaires du modèle et de regrouper les couples $(\hat{\beta}_\ell, \hat{\theta}_\ell)$ selon le voisinage auquel $\hat{\theta}_\ell$ appartient. On peut ensuite se donner les risques d'estimation suivants:

$$\sum_{k=1}^{s} \Big| \beta_k^\star - \sum_{\ell \in S_k(r)} \hat{\beta}_\ell \Big| \qquad \text{et} \qquad \sum_{\ell \in S(r)^c} |\hat{\beta}_\ell|, \tag{1.4}$$

où,

- l'ensemble d'indices $S(r)$ donné par:

$$S(r) = \bigcup_{1\leq k\leq s} S_k(r) \quad \text{avec} \quad S_k(r) = \Big\{\ell, \quad \hat{\beta}_\ell \neq 0 \quad \text{et} \quad \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star) \leq r\Big\} \tag{1.5}$$

  correspondant à l'ensemble des indices ℓ tels que le paramètre actif $\hat{\theta}_\ell$ est proche d'un des vrais paramètres du modèle $\theta_k^\star$;

- l'ensemble d'indices $S^c(r)$ est le complémentaire de $S(r)$ et correspond aux indices associés à des paramètres estimés loin de ceux du modèle.

On peut remarquer que les quantités présentées dans (1.4) ne nous prémunissent pas contre le cas problématique d'une compensation entre les estimateurs $\hat{\beta}_\ell$ approchant le même coefficient linéaire $\beta_k^\star$. Typiquement, on voudrait pouvoir éviter le cas où de grandes valeurs de coefficients estimés $\hat{\beta}_\ell$ de signes différents s'additionnent pour approcher un petit (en valeur absolue) coefficient linéaire $\beta_k^\star$. Ainsi, nous introduisons un troisième risque d'estimation:

$$\sum_{k=1}^{s} \Big| |\beta_k^\star| - \sum_{\ell\in S_k(r)} |\hat{\beta}_\ell| \Big|. \tag{1.6}$$

Pour quantifier la qualité de la reconstruction d'un signal, il est naturel de considérer le risque de prédiction:

$$\Big\| \beta^\star \Phi_T(\vartheta^\star) - \hat{\beta}\, \Phi_T(\hat{\vartheta}) \Big\|_T, \tag{1.7}$$

où pour $s \in \mathbb{N}$ et $\vartheta \in \Theta^s$, $\Phi_T(\vartheta) \in H_T^s$ est défini par le vecteur colonne:

$$\Phi_T(\vartheta) = (\phi_T(\theta_1), \ldots, \phi_T(\theta_s))^\top,$$

avec la convention que $\beta\Phi_T(\vartheta) = 0$ pour $s = 0$. Ce risque mesure la proximité entre le signal original et le signal reconstruit par les estimateurs.

L'extension de ce risque de prédiction pour la reconstruction simultanée de plusieurs signaux (éventuellement un continuum) s'écrit:

$$\Big\| B^\star\, \Phi_T(\vartheta^\star) - \hat{B}\, \Phi_T(\hat{\vartheta}) \Big\|_{L_T} = \sqrt{\int_{\mathcal{Z}} \Big\| B^\star(z)\, \Phi_T(\vartheta^\star) - \hat{B}(z)\, \Phi_T(\hat{\vartheta}) \Big\|_T^2 \, \nu(\mathrm{d}z)}. \tag{1.8}$$

Dans cette thèse, nous tâcherons de donner à toutes ces quantités des bornes valables en grande probabilité en fonction de la parcimonie du modèle et de la qualité des observations.

1.4.1 Vers des méthodes sans grille

Rappelons que les composantes sous-jacentes aux signaux considérés sont issues d'une famille continûment paramétrée sur l'ensemble Θ. En vue de construire des estimateurs pour les coefficients linéaires et les paramètres des modèles (1.1) et (1.3), une première approche consiste à discrétiser l'espace des paramètres Θ et à approcher le signal par une combinaison linéaire de composantes dont les paramètres se trouvent dans la grille de discrétisation (voir [Tang et al., 2013a]). Ce faisant, l'estimation des signaux est réduite à l'estimation de coefficients linéaires. Malheureusement, si de tels estimateurs sont attrayants du fait de la simplicité de leur définition, la discrétisation conduit à des problèmes aussi bien numériques que théoriques. En conséquence, il est souvent préférable d'utiliser des estimateurs ne nécessitant pas de grille sur l'espace des paramètres. Notons toutefois que l'étude des estimateurs linéaires parcimonieux associés à une grille fournit de précieux outils pour le développement de méthodes sans grille.

Estimateurs parcimonieux

Supposons que l'espace des paramètres Θ soit discrétisé sur une grille de K points $\vartheta^{\mathcal{G}} = (\theta_1^{\mathcal{G}}, \cdots, \theta_K^{\mathcal{G}})$. Nous pourrions être tentés d'approcher le signal par une combinaison linéaire de composantes appartenant au dictionnaire fini $(\phi_T(\theta_k^{\mathcal{G}}), 1 \leq k \leq K)$. Bien sûr, nous aimerions construire des estimateurs proches des vrais coefficients et paramètres du modèle. En particulier, nous voudrions éviter d'approcher le signal par un mélange contenant un nombre de composantes actives beaucoup plus grand que s. En somme, nous recherchons une représentation parcimonieuse du signal dans un dictionnaire fini. Trouver la représentation la plus parcimonieuse d'un signal dans un dictionnaire fini peut se faire en minimisant sous contraintes une pseudo-norme ℓ_0 comptant le nombre d'entrées non nulles d'un vecteur. Il s'agit d'un problème NP-difficile (voir [Natarajan, 1995]) devenant insoluble numériquement lorsque la taille K du dictionnaire $(\phi_T(\theta_k^{\mathcal{G}}), 1 \leq k \leq K)$ est grande. Par conséquent, l'idée de minimiser une norme ℓ_1, additionnant les valeurs absolues des entrées d'un vecteur, plutôt que la pseudo-norme ℓ_0, a fait florès dans la littérature. Cette relaxation convexe du précédent problème s'avère beaucoup plus accessible. Elle a été utilisée sous sa forme lagrangienne en géophysique par [Santosa and Symes, 1986], étudiée ensuite dans la communauté de l'acquisition comprimée par [Chen and Donoho, 1994], [Chen et al., 1998], [Donoho et al., 2006] sous le nom de *basis pursuit denoising*, et dans la communauté statistique par [Tibshirani, 1996] sous le nom de Lasso (pour *Least absolute shrinkage and selection operator*). Sous sa forme lagrangienne, le problème consiste à minimiser la somme d'un terme d'adéquation aux données (garantissant que le mélange approximant est proche de l'observation) et d'une pénalité ℓ_1 pondérée (assurant que la solution promue est parcimonieuse). Le problème s'écrit pour un paramètre de pénalisation $\kappa > 0$:

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^K}{\operatorname{argmin}} \quad \frac{1}{2}\|y - \beta\Phi_{\mathcal{G}}\|_T^2 + \kappa\|\beta\|_{\ell_1}, \tag{1.9}$$

où $\Phi_{\mathcal{G}} = (\phi_T(\theta_1^{\mathcal{G}}), \cdots, \phi_T(\theta_K^{\mathcal{G}}))^\top$ et $\|\beta\|_{\ell_1} = \sum_{k=1}^K |\beta_k|$. Des algorithmes et des solveurs efficaces sont disponibles pour trouver une solution au problème d'optimisation (1.9) (voir par exemple l'algorithme FISTA dans [Beck and Teboulle, 2009]). De plus, il est montré dans [Donoho, 2006] que la solution du Lasso est une bonne approximation de la représentation la plus parcimonieuse du signal dans le dictionnaire fini $\{\phi_T(\theta_1^{\mathcal{G}}), \cdots, \phi_T(\theta_K^{\mathcal{G}})\}$. D'autres estimateurs faciles à mettre en œuvre peuvent être trouvés dans la littérature, comme le *Dantzig selector* de [Candès and Tao, 2007]. Ces estimateurs ont été largement étudiés par la communauté statistique. En particulier, des bornes de prédiction et d'estimation en fonction de la parcimonie s, de la taille du dictionnaire K et de la qualité des observations T, valables en grande probabilité, ont été établies pour ceux-ci. Nous renvoyons à [Bunea et al., 2007], [Bickel et al., 2009] et [Koltchinskii, 2009] pour les énoncés précis et les preuves de ces résultats. Des bornes portant sur l'espérance des risques d'estimation et de prédiction ont également été montrées pour le Lasso dans [Bellec and Tsybakov, 2017]. Ces bornes sont établies sous des hypothèses de cohérence sur le dictionnaire. De nombreuses variantes de ces hypothèses de cohérence peuvent être trouvées dans la littérature. Parmi celles-ci, on trouve la propriété d'isométrie restreinte (RIP), la condition de valeur propre restreinte (RE) ou encore les conditions de compatibilité. Mentionnons le travail de [van de Geer, 2016] pour un aperçu de ces conditions. Soulignons également que les bornes susmentionnées sont des conséquences d'inégalités oracles. Ces dernières quantifient la performance d'un estimateur en le comparant à un estimateur idéal reposant sur une information parfaite du modèle (par exemple, la connaissance des coefficients linéaires à estimer). Nous renvoyons à [Candès, 2006] et [van de Geer, 2016] pour une présentation complète de cet outil.

Lorsqu'il s'agit de quantifier la performance d'un estimateur, le paradigme minimax se révèle pertinent. Un estimateur est dit minimax pour un risque donné si son risque maximal est minimal en comparaison avec tous les estimateurs possibles. En somme, l'estimateur minimax obtient les meilleures performances dans le pire cas possible autorisé par le problème.

Ainsi, il est d'usage en statistiques de donner des bornes inférieures à l'infinimum sur les estimateurs, du supremum sur les paramètres à estimer, des risques d'estimation et de prédiction. Dans le cas de l'estimateur Lasso, les bornes supérieures sur les risques de prédiction et d'estimation correspondent quasiment aux bornes inférieures (se référer à [Candès and Davenport, 2013] et [Raskutti et al., 2011]). Ainsi, l'estimateur Lasso est quasiment minimax.

Estimateurs parcimonieux par groupe

Lorsqu'on dispose d'un ensemble d'observations partageant une structure commune, comme dans le modèle (1.3), il est naturel de penser qu'estimer les signaux séparément n'est peut-être pas la meilleure façon de procéder. Dans le modèle (1.3), la parcimonie du jeu de données correspond au cardinal de l'union des composantes actives des signaux. On pourrait à nouveau être tenté de discrétiser l'espace des paramètres Θ pour utiliser les estimateurs parcimonieux par groupe disponibles dans la littérature pour les modèles linéaires de grande dimension. Ces estimateurs promeuvent les approximations utilisant un petit nombre de composantes pour approcher l'ensemble des données. Supposons que $\mathcal{Z}$ soit fini de cardinal n et que $H_T = \mathbb{R}^T$, de sorte que l'espace d'observation $L_T = L^2(\nu, H_T)$ soit identifié à $\mathbb{R}^{n\times T}$. Dans cet exemple, n signaux de T points sont observés. L'estimateur group-Lasso, introduit dans [Yuan and Lin, 2006], est particulièrement adapté à ce contexte. De nombreux travaux peuvent être trouvés dans la littérature statistique sur cet estimateur (voir [Huang and Zhang, 2010], [Nardi and Rinaldo, 2008], [Bühlmann and van de Geer, 2011], [Obozinski et al., 2011], [Lounici et al., 2011] et les références mentionnées dans ces papiers). Le problème d'optimisation group-Lasso s'écrit pour un paramètre de pénalisation $\kappa > 0$:

$$\hat{B} \in \operatorname*{argmin}_{B\in\mathbb{R}^{n\times K}} \quad \frac{1}{2}\|Y - B\Phi_{\mathcal{G}}\|^2_{L_T} + \kappa\|B\|_{\ell_1,\ell_2}, \tag{1.10}$$

où $\|B\|_{\ell_1,\ell_2} = \sum_{k=1}^K \|B_{\cdot,k}\|_{\ell_2}$ avec $\|\cdot\|_{\ell_2}$ la norme euclidienne et $B_{\cdot,k}$ désigne la k-ième colonne de la matrice B. Dans [Lounici et al., 2011], il est montré que cet estimateur tire profit de la parcimonie de groupe et surpasse l'estimateur Lasso dans certains cas. Les auteurs prouvent, sous des conditions de compatibilité, des bornes supérieures pour les risques d'estimation et de prédiction qui sont optimales (à un facteur logarithmique près) au sens minimax.

Inconvénients liés à la discrétisation de l'espace des paramètres

Les estimateurs parcimonieux et parcimonieux par groupe obtenus après discrétisation de l'espace des paramètres Θ se heurtent à des problèmes numériques et théoriques. Définissons l'ensemble $\hat{S} = \{k,\, \hat{\beta}_k \neq 0\}$ contenant les indices des entrées non nulles de la solution $\hat{\beta}$ de (1.9), ainsi que le risque:

$$\left\|\beta^\star\, \Phi_T(\vartheta^\star) - \hat{\beta}\, \Phi_T(\vartheta^{\mathcal{G}})\right\|_T, \quad \text{avec la grille} \quad \vartheta^{\mathcal{G}} = (\theta_1^{\mathcal{G}}, \cdots, \theta_K^{\mathcal{G}}). \tag{1.11}$$

En résolvant le problème Lasso (1.9), nous souhaiterions que la quantité (1.11) tende vers zéro et que les estimateurs $(\hat{\beta}_{\hat{S}}, \vartheta^{\mathcal{G}}_{\hat{S}})$ tendent vers $(\beta^\star, \vartheta^\star)$ (à une permutation près) lorsque la qualité de l'information augmente, c'est-à-dire lorsque T devient grand. Il est clair qu'il n'y a *a priori* aucune raison pour que cela se produise à moins que les paramètres à estimer soient inclus dans la grille, *i.e.* que nous ayons l'inclusion $\{\theta_1^\star, \cdots, \theta_s^\star\} \subset \{\theta_1^{\mathcal{G}}, \cdots, \theta_K^{\mathcal{G}}\}$. Par conséquent, il faudrait raffiner la grille sur l'espace des paramètres Θ. Malheureusement, cela soulève de nouveaux problèmes. Premièrement, quand l'espace des paramètres est de grande dimension, la discrétisation devient extrêmement coûteuse. Ensuite, même lorsque l'espace des paramètres est unidimensionnel, le raffinement de la grille induit de fortes corrélations entre les composantes du dictionnaire, conduisant à des instabilités numériques. Les récents travaux de [Duval and Peyré, 2017a] et [Duval and Peyré, 2017b] fournissent un

autre argument en défaveur des méthodes nécessitant une grille. Dans ces articles, les auteurs considèrent des combinaisons linéaires de pics continûment paramétrés par leur position. Ils montrent que la résolution du problème Lasso (1.9) sur une grille fine conduit à sur-estimer le nombre de pics dans le mélange. Des clusters de pics sont obtenus autour des pics à estimer. Une amélioration de cette méthode a été proposée dans [Ekanadham et al., 2011] sous le nom de *continuous basis pursuit*. Le problème d'optimisation qui y est décrit vise à limiter l'effet de la grille en utilisant une approximation de Taylor au premier ordre des pics. Néanmoins, dans [Duval and Peyré, 2017b], il est démontré que cette amélioration du Lasso engendre également des clusters de pics lorsque la grille devient fine. Pour toutes ces raisons, nous nous concentrerons dans cette thèse sur des méthodes ne nécessitant pas de grille sur l'espace des paramètres.

1.4.2 Méthodes sans grille adaptées aux dictionnaires continus

Lorsqu'une borne K sur le nombre s de composantes dans le mélange est connue, on peut, comme proposé dans [Golub and Pereyra, 1973] et [Kaufman, 1975], décomposer un signal dans un dictionnaire continu en résolvant le problème des moindres carrés non linéaire suivant:

$$\min_{\beta\in\mathbb{R}^K,\vartheta\in\Theta^K} \quad \|y-\beta\,\Phi_T(\vartheta)\|_T^2. \tag{1.12}$$

Cependant, en présence de bruit, les solutions à ce type de problème dépendent fortement de K; ce qui n'est pas satisfaisant. En particulier, cette méthode a tendance à largement sur-estimer le nombre de composantes dans le mélange lorsque K est grand.

De nombreux travaux de recherche ont visé à reconstruire des mélanges parcimonieux de composantes issues d'un dictionnaire continu sans discrétiser l'espace des paramètres. Mentionnons les travaux pionniers de [Mallat and Zhang, 1993] qui ont introduit l'algorithme de *Matching Pursuit* (MP) dans le but de décomposer un signal dans un dictionnaire généré par une fonction continûment translatée, dilatée et modulée. Une contribution remarquable provient également de l'article [de Castro and Gamboa, 2012] dans lequel une extension convexe et continue du problème Lasso, appelée Beurling Lasso (ou BLasso), est formulée. Il s'agit d'un problème de minimisation sur un espace de mesure qui, lorsque ses solutions sont discrètes, fournit des estimateurs aux paramètres du modèle (1.1). Néanmoins, pour certains modèles que nous considérerons, les solutions du BLasso ne sont *a priori* pas discrètes. Nous étudierons donc un problème un peu différent nécessitant la connaissance d'une borne K sur le nombre de composantes s dans le mélange.

Estimateurs issus d'un problème pénalisé

Afin de palier aux problèmes rencontrés dans l'étude théorique et la résolution numérique de (1.12), nous formulons un problème pénalisé dans la lignée des travaux sur les modèles linéaires parcimonieux. Ce problème s'écrit pour un paramètre de pénalisation $\kappa>0$ et une borne K sur la parcimonie s:

$$(\hat\beta,\hat\vartheta)\in \operatorname*{argmin}_{\beta\in\mathbb{R}^K,\vartheta\in\Theta_T^K} \quad \frac{1}{2}\|y-\beta\,\Phi_T(\vartheta)\|_T^2+\kappa\|\beta\|_{\ell_1}, \tag{1.13}$$

où Θ_T est un sous-ensemble compact de Θ. Puisque Θ_T est compact, ce problème admet toujours une solution.

Exemple 1.5 (Processus à temps discret sur une grille régulière). Poursuivons l'exemple 1.2. Le support de la mesure λ_T est compris dans l'intervalle $[a_T,b_T]$. Supposons que les composantes du dictionnaire soient des pics continûment translatés sur $\mathbb{R}$. Il est alors légitime de rechercher les paramètres de localisation sur un sous-ensemble plus petit que la fenêtre d'observation $[a_T,b_T]$, et ainsi restreindre l'optimisation (1.13) à l'ensemble compact:

$$\Theta_T=[(1-\epsilon)a_T,(1-\epsilon)b_T]\subset[a_T,b_T] \quad \text{avec un rétrécissement donné } \epsilon>0.$$

Exemple 1.6 (Processus à temps continu sur le tore). Poursuivons l'exemple 1.3. Supposons que les composantes du dictionnaire soient des pics continûment translatés sur le tore. Ce dernier est compact et on peut réaliser l'optimisation (1.13) sur $\Theta = \mathbb{R}/\mathbb{Z}$.

Nous utiliserons de nombreux outils théoriques du BLasso généralisant le problème (1.13). Le BLasso consiste sous sa forme lagrangienne à minimiser sur un espace de mesures une fonction objectif composée d'un terme d'adéquation aux données et d'une pénalité pondérée. Il s'écrit pour un paramètre de pénalisation $\kappa > 0$:

$$\min_{\mu \in \mathcal{M}(\Theta_T)} \quad \frac{1}{2} \|y - \langle \phi_T, \mu \rangle\|_T^2 + \kappa \|\mu\|_{TV}, \tag{1.14}$$

où $\mathcal{M}(\Theta_T)$ désigne l'ensemble des mesures de Radon sur l'espace des paramètres Θ_T, $\|\cdot\|_{TV}$ est la norme en variation totale sur les mesures et $\langle \phi_T, \mu \rangle = \int \phi_T(\theta)\, \mu(\mathrm{d}\theta)$. En prenant $\mu = \sum_{k=1}^K \beta_k \, \delta_{\theta_k}$ où δ_x désigne une mesure de Dirac localisée en x, les fonctions objectifs des problèmes (1.14) et (1.13) sont égales. En résolvant (1.14), on souhaite récupérer une mesure parcimonieuse, c'est-à-dire une somme d'un petit nombre de mesures de Dirac (ou atomes) pondérées. Dans ce cas, les amplitudes et les emplacements des atomes estiment respectivement les coefficients linéaires dans le mélange et les paramètres des composantes. Depuis son introduction, le BLasso a suscité un grand intérêt dans les communautés de l'acquisition comprimée et des statistiques. Citons les travaux précurseurs de [Candès and Fernandez-Granda, 2014] et [Candès and Fernandez-Granda, 2013] dans lesquels le BLasso est résolu dans le cadre de la super-résolution par une méthode d'optimisation semi-définie positive (SDP), ainsi que l'article [Bredies and Pikkarainen, 2013] dans lequel une variante de l'algorithme de Frank-Wolfe est proposée pour le résoudre dans un cadre général. Il a ensuite été utilisé, entre autres, pour la déconvolution de pics dans [Duval and Peyré, 2015] ainsi que pour des modèles de mélange de densité dans [De Castro et al., 2021]. Mentionnons également les travaux de [Tang et al., 2013b] et [Tang et al., 2015] autour d'une formulation d'un problème d'optimisation proche du BLasso appelé *atomic norm minimization*.

S'il est aisé de montrer que le problème convexe (1.14) admet une solution, celle-ci n'est pas *a priori* composée d'un nombre fini d'atomes. Or, les solutions du BLasso sont difficilement interprétables lorsqu'elles ne sont pas atomiques. Typiquement, on ne peut pas directement définir d'estimateurs pour les paramètres du modèle (1.1). Lorsque l'espace $H_T = \mathbb{R}^T$, il existe d'après [Boyer et al., 2019] une solution atomique au BLasso composée d'au plus T mesures de Dirac. Ainsi, lorsque la borne K est supérieure ou égale à T, cette solution minimise le problème restreint (1.13). Inversement, les solutions de (1.13) sont des solutions du problème BLasso. À notre connaissance, il n'existe pas de tel résultat lorsque l'espace H_T est un espace de Hilbert quelconque; ce qui correspond au cadre de cette thèse. De ce fait nous nous intéresserons à la restriction (1.13) du problème BLasso aux mesures atomiques composées d'au plus K atomes, où K est une borne arbitrairement grande sur la parcimonie s.

Généralisation à la parcimonie de groupe

Afin d'estimer les paramètres du modèle (1.3) dans lequel plusieurs signaux sont observés (éventuellement un continuum), on pourrait naïvement procéder à des estimations individuelles en résolvant (1.13) pour toute observation $Y(z)$ indexée par un élément z de l'ensemble $\mathcal{Z}$. Néanmoins, nous souhaitons tirer profit de la structure commune des différents signaux. Dans la lignée des travaux sur les modèles linéaires parcimonieux par groupe, on formule le problème d'optimisation suivant avec un paramètre de pénalisation $\kappa > 0$, une borne K sur la parcimonie de groupe s et $p \in [1, 2]$:

$$(\hat{B}, \hat{\vartheta}) \in \operatorname*{argmin}_{B \in L^2(\nu, \mathbb{R}^K), \vartheta \in \Theta_T^K} \quad \frac{1}{2\nu(\mathcal{Z})} \|Y - B\Phi_T(\vartheta)\|_{L_T}^2 + \kappa \|B\|_{\ell_1, L^p(\nu)}, \tag{1.15}$$

où pour $z \mapsto B(z) = (B_1(z), \dots, B_K(z))$ dans $L^2(\nu, \mathbb{R}^K)$:

$$\|B\|_{\ell_1, L^p(\nu)} = \sum_{k=1}^{K} \|B_k\|_{L^p(\nu)}.$$

Le terme d'adéquation aux données correspond à l'intégration sur l'espace $(\mathcal{Z}, \mathcal{F}, \nu)$ (indexant les signaux observés) de celui du problème (1.13). Quant à la pénalisation, elle correspond à la somme des énergies de chaque composante à travers la collection de signaux; l'énergie d'une composante étant exprimée par une norme $L^p(\nu)$ sur les coefficients linéaires associés. Lorsque l'ensemble $\mathcal{Z}$ est fini de cardinal n, l'espace des coefficients linéaires est identifié à $\mathbb{R}^{n \times K}$. En prenant $p = 2$, la pénalisation devient celle de l'estimateur group-Lasso introduit pour les modèles linéaires, voir (1.10). Lorsque que $p = 1$, la pénalisation est celle du Lasso utilisée pour les modèles de régression linéaires multiples et ne favorise pas la parcimonie de groupe. En permettant à p de prendre des valeurs entre 1 et 2, la parcimonie des solutions peut être réglée.

Notons qu'une généralisation du problème BLasso a été proposée dans [Golbabaee and Poon, 2022] pour l'estimation simultanée de n signaux. Appliquée à des mesures atomiques, la pénalité utilisée dans leur problème de minimisation est une combinaison convexe des pénalités Lasso et group-Lasso. Les solutions de cette généralisation du BLasso ne sont *a priori* pas des mesures atomiques lorsque H_T est de dimension infinie. C'est la raison pour laquelle nous étudions le problème (1.15). En effet, lorsque $p \in (1, 2]$, le problème (1.15) admet toujours une solution. En outre si H_T est de dimension finie, il admet une solution pour $p \in [1, 2]$.

1.6.1 Bornes d'estimation et de prédiction

Nous donnerons dans cette thèse des bornes d'estimation et de prédiction valables en grande probabilité pour les estimateurs issus du problème (1.13) et de sa généralisation (1.15). Ces bornes reposeront sur des fonctions interpolatrices, appelées certificats, ainsi que sur le contrôle de la probabilité qu'un processus gaussien régulier dépasse un certain niveau sur un intervalle donné.

Les certificats

Les certificats sont des outils essentiels dans l'étude du BLasso. En particulier, ils permettent d'établir des résultats d'identifiabilité et de robustesse au bruit (se référer à [de Castro and Gamboa, 2012], [Candès and Fernandez-Granda, 2013], [Candès and Fernandez-Granda, 2014], [Duval and Peyré, 2015], [Poon et al., 2021]). Les certificats sont des fonctions à valeurs réelles, définies sur Θ, de la forme:

$$\eta : \theta \mapsto \langle \phi_T(\theta), p \rangle_T \,,$$

où $\langle \cdot, \cdot \rangle_T$ désigne le produit scalaire associé à H_T et p est un élément de H_T déterminé de façon à ce que la fonction η interpole -1 ou 1 aux points $\theta_k^\star$, ait une dérivée seconde non nulle en ces points et soit strictement comprise entre -1 et 1 partout ailleurs. Dans [Poon et al., 2021], la construction de certificats est étendue à une large gamme de dictionnaires. L'article couvre le cas où l'espace des paramètres est multidimensionnel et où le dictionnaire n'est pas forcément généré par une fonction continûment translatée. En particulier, dans le cas où Θ est unidimensionnel de telles fonctions interpolatrices existent dès lors que les paramètres non-linéaires du signal sont suffisamment séparés deux à deux au sens de la métrique riemannienne définie pour $\theta, \theta' \in \Theta$ par:

$$\mathfrak{d}_T(\theta, \theta') = \inf_{\gamma} \int_0^1 |\dot{\gamma}_s| \sqrt{\partial_{x,y} \mathcal{K}_T(\gamma_s, \gamma_s)} \, \mathrm{d}s$$

où l'infinimum porte sur l'ensemble des chemins réguliers $\gamma : [0,1] \to \Theta$ tels que $\gamma_0 = \theta$ et $\gamma_1 = \theta'$ et $\partial_{x,y}\mathcal{K}_T$ désigne la dérivée selon la première et la deuxième variable du noyau $\mathcal{K}_T$ défini par:

$$\mathcal{K}_T(\theta,\theta') \mapsto \langle \phi_T(\theta), \phi_T(\theta') \rangle_T \quad \text{avec} \quad \phi_T \quad \text{défini par (1.2)}.$$

La distance $\mathfrak{d}_T$ ainsi définie est invariante par reparamétrisation de l'espace Θ. En effet, les métriques riemanniennes $\mathfrak{d}_T$ et $\mathfrak{d}_T^{\mathrm{h}}$ associées respectivement au noyau $\mathcal{K}_T(\cdot,\cdot)$ et au noyau déformé $\mathcal{K}_T^{\mathrm{h}} = \mathcal{K}_T(\mathrm{h}(\cdot),\mathrm{h}(\cdot))$ où h est un difféomorphisme suffisamment régulier, satisfont l'égalité : $\mathfrak{d}_T(\theta,\theta') = \mathfrak{d}_T^{\mathrm{h}}(\mathrm{h}^{-1}(\theta),\mathrm{h}^{-1}(\theta'))$. Cette propriété d'invariance est partagée avec la fonction objectif du problème d'optimisation considéré et motive l'utilisation de la distance $\mathfrak{d}_T$ sur l'espace des paramètres.

Bornes valables en grande probabilité

Dans [Duval and Peyré, 2015], il est montré, à condition que les quantités $\|w_T\|_T/\kappa$ et κ soient inférieures à un certain seuil et sous une hypothèse d'existence de certificats, que la solution du BLasso est une combinaison linéaire d'exactement s atomes dont les positions et les amplitudes approchent les coefficients linéaires et les paramètres non-linéaires du signal à une perturbation près, de l'ordre de $\|w_T\|_T$ (voir [Duval and Peyré, 2015, Theorem 2]). Dans [Golbabaee and Poon, 2022] ce résultat est étendu au group-BLasso. Malheureusement, pour de nombreux modèles que l'on considérera la condition requise sur $\|w_T\|_T/\kappa$ n'est pas vérifiée. En outre, ces bornes se révèlent sous optimales dans les régimes de bruit étudiés. Prenons par exemple $H_T = \mathbb{R}^T$ muni du produit scalaire usuel, noté ici $\langle \cdot,\cdot \rangle_T$, et supposons que w_T est un vecteur Gaussien avec des entrées indépendantes de variance 1. Il est aisé de voir que la variable aléatoire $\|w_T\|_T$ est alors distribuée comme la racine carrée d'une variable du χ^2 d'ordre T de sorte que $\lim_{T\to+\infty} \mathbb{E}[\|w_T\|_T] = +\infty$. Dans [Poon et al., 2021], des bornes sur des quantités similaires à (1.4) sont données dans un cadre très général. Seulement, celles-ci dépendent linéairement de $\|w_T\|_T$. Pour certains modèles que nous considérerons, ces bornes sont sous-optimales. Nous tâcherons dans cette thèse d'employer des méthodes différentes pour contrôler avec une grande probabilité les risques de prédiction (1.7) et (1.8) ainsi que les risques d'estimation (1.4) et (1.6) en fonction de la qualité des observations et en particulier du niveau du bruit exprimé par:

$$\sup_{f\in\mathcal{F}} \quad \mathrm{Var}\,\langle f, w_T \rangle_T\,,$$

où $\mathcal{F}$ est une famille d'éléments de H_T liés au dictionnaire. Des bornes quasiment optimales au sens minimax ont été établies sur ces risques dans [Tang et al., 2015] et [Boyer et al., 2017] dans un contexte particulier de la super-résolution. Ces résultats reposent sur deux ingrédients fondamentaux. D'abord, les certificats issus de [Candès and Fernandez-Granda, 2013] sont utilisés. Ensuite, les bornes établies font apparaître des suprema de processus gaussiens de la forme:

$$\sup_{\theta\in\Theta} \left\langle \partial_\theta^i \phi_T(\theta), w_T \right\rangle_T, \quad \text{avec} \quad i = 0,1,2,$$

dont les queues de distribution sont contrôlées à l'aide d'inégalités dérivées de formules de type Rice. Introduites dans [Rice, 1944] pour des processus gaussiens réguliers et stationnaires, les formules de Rice fournissent les moments de la variable aléatoire N_u comptant le nombre de fois où un processus prend la valeur u sur un intervalle donné. Elles permettent, en particulier, de contrôler la probabilité qu'un processus régulier dépasse un certain niveau $u > 0$. Nous renvoyons aux ouvrages [Adler and Taylor, 2007] et [Azaïs and Wschebor, 2009] pour une présentation détaillée de ces formules.

1.6.2 Aspects numériques

Nous avons déjà évoqué les travaux de [Candès and Fernandez-Granda, 2014] utilisant une méthode d'optimisation semi-définie positive afin de résoudre le BLasso dans le cadre de la super-résolution. Il s'avère que dans le cadre général, des modifications du classique algorithme de Frank-Wolfe sont efficaces pour le résoudre. Citons notamment les travaux de [Bredies and Pikkarainen, 2013], [Boyd et al., 2017], [Denoyelle et al., 2020] et [Golbabaee and Poon, 2022]. Mentionnons qu'une autre méthode sans grille, appelée Conic Particle Gradient Descent (CPGD), a été développée pendant l'élaboration de cette thèse (se référer à [Chizat, 2021]). Ces algorithmes recherchent des solutions atomiques aux problèmes BLasso et group-BLasso. Ainsi, ils sont applicables aux problèmes restreints introduits dans (1.13) et (1.15).

1.7 Procédures de test

Estimer les paramètres du modèle (1.1) revient à déterminer la distribution de probabilité ayant généré l'observation y. Lorsqu'il s'agit plutôt de tester si cette distribution satisfait certaines propriétés, il n'est pas forcément nécessaire d'estimer tous les paramètres du modèle. Cela fait du test statistique un domaine complémentaire à l'estimation.

L'analyse de procédures de test est ici motivée par l'étude du modèle de déconvolution de pics et plus particulièrement par ses applications en spectroscopie. Nous souhaiterions décider, à partir d'une observation y issue du modèle (1.1), si le signal inconnu $\beta^\star \Phi_T(\vartheta^\star)$, avec $\beta^\star \in (\mathbb{R}^*)^s$ et $\vartheta^\star = (\theta_1^\star, \cdots, \theta_s^\star) \in \Theta^s$, est égal à un signal de référence $\beta^0 \Phi_T(\vartheta^0)$ pour certains vecteurs connus $\beta^0 \in (\mathbb{R}^*)^{s^0}$ et $\vartheta^0 = (\theta_1^0, \cdots, \theta_s^0) \in \Theta^{s^0}$. Plus formellement, nous souhaiterions distinguer une hypothèse nulle H_0 d'une hypothèse alternative H_1 définies comme suit:

$$\begin{cases} H_0 : & \beta^\star \Phi_T(\vartheta^\star) = \beta^0 \Phi_T(\vartheta^0), \\ H_1(\rho) : & \left\| \beta^\star \Phi_T(\vartheta^\star) - \beta^0 \Phi_T(\vartheta^0) \right\|_T \geq \rho, \end{cases}$$

où $\rho > 0$ est un paramètre de séparation. Une procédure de test visant à distinguer H_0 de H_1 est une fonction mesurable de l'observation y à valeurs dans $\{0, 1\}$ valant 0 lorsque l'hypothèse nulle est acceptée et 1 lorsque celle-ci est rejetée. Le risque associé à une procédure de test correspond alors à la somme des probabilités de rejeter H_0 à tort et d'accepter H_0 à tort. L'enjeu est de construire des procédures dont le risque est inférieur à un certain niveau $\alpha \in (0, 1)$, tout en permettant à la séparation ρ d'être la plus petite possible. Nous renvoyons à l'ouvrage [Ingster and Suslina, 2003] pour une présentation complète des tests d'adéquation à un modèle. Notons que les hypothèses présentées ci-dessus couvrent, en fixant $s^0 = 0$, le cadre de la détection de signal. En effet, pourvu que la matrice symétrique $(\langle \phi_T(\theta_k^\star), \phi_T(\theta_\ell^\star) \rangle_T)_{1 \leq k, \ell \leq s}$ ait des valeurs propres encadrées par deux constantes strictement positives, le test précédent permet de distinguer, pour un paramètre de séparation ρ' du même ordre que ρ, les hypothèses:

$$\begin{cases} H_0 : & s = 0 \\ H_1(\rho) : & \|\beta^\star\|_{\ell_2} \geq \rho', \end{cases} \tag{1.16}$$

où $\|\beta^\star\|_{\ell_2}$ est l'énergie du signal exprimée par la norme euclidienne des amplitudes des pics. Dès lors, la question est de déterminer l'énergie minimale permettant la détection d'un signal en présence de bruit.

Un deuxième problème de test auquel nous nous intéresserons consiste à décider si le signal observé est une combinaison linéaire de pics situés dans une liste prescrite d'emplacements ou s'il contient des pics supplémentaires. En somme, cela revient à tester si l'ensemble des emplacements des pics $\mathcal{Q}^\star$ est inclus dans un sous-ensemble fini et connu $\mathcal{Q}^0$ de Θ de cardinal s^0. Cette configuration est issue d'une application à la spectroscopie (voir [Butucea et al.,

2021]), où la présence de composants chimiques supplémentaires à ceux prescrits indique un vieillissement ou des modifications substantielles du matériau analysé.

1.8 Contributions

Cette thèse aborde des problèmes d'estimation et de test pour des mélanges de composantes parcimonieux provenant d'un dictionnaire continûment paramétré sur l'espace unidimensionnel Θ. Une grande variété de modèles de régression non linéaires sont considérés dans un cadre unifié. Nous considérons un large panel de dictionnaires continus, d'espaces d'observations et de bruits additifs gaussiens.

La thèse s'articule autour de quatre chapitres. Dans le chapitre 2, nous traitons le cas où un seul élément aléatoire de l'espace de Hilbert H_T est observé. Nous établissons des bornes valables en grande probabilité pour les risques d'estimation et de prédiction (1.4), (1.6) et (1.7) associés aux estimateurs définis par le problème d'optimisation pénalisé (1.13). Ces bornes sont illustrées par l'exemple du modèle de déconvolution de pics gaussiens. Le chapitre 3 aborde le cas plus général où un ensemble d'éléments de H_T est observé. Des bornes de prédiction sont données sur le risque (1.8) avec une certaine probabilité. Cette probabilité est explicitement minorée en fonction des données du problème lorsque le nombre d'observations est fini. Le chapitre 4 porte sur des modèles de translation dans lesquels le dictionnaire dépend d'un paramètre d'échelle. Nous montrons que les bornes valables en grande probabilité fournies dans le chapitre 2 peuvent être établies sous des contraintes moins restrictives sur les paramètres non-linéaires du mélange. Des procédures de test sont également analysées. Enfin, le chapitre 5 consiste en une application numérique du problème d'estimation des paramètres du modèle de déconvolution de pics gaussiens. Cette application est motivée par l'étude du vieillissement de revêtements en polychoroprène via la spectroscopie infrarouge. Elle a donné lieu à l'implémentation d'une méthode numérique en Python pour Électricité de France.

1.8.1 Contributions du chapitre 2

Le contenu de ce chapitre est issu de [Butucea et al., 2022a].

Le résultat principal du chapitre 2 donne des bornes en grande probabilité pour le risque de prédiction (1.7) et les risques d'estimation (1.4) et (1.6) associés aux estimateurs $\hat{\beta}$ et $\hat{\vartheta}$ des paramètres $\beta^\star$ et $\vartheta^\star$, obtenus en résolvant le problème (1.13) pour un paramètre de pénalisation κ et pour une borne K sur la parcimonie.

On mesure la colinéarité entre deux éléments du dictionnaire continu à l'aide du noyau $\mathcal{K}_T(\theta,\theta') = \langle \phi_T(\theta), \phi_T(\theta')\rangle_T$ défini sur Θ^2. Des propriétés de régularité et de concavité locale sur la diagonale sont requises sur celui-ci. En pratique, il peut-être difficile de vérifier ces propriétés. De ce fait, nous considérons un noyau approximant $\mathcal{K}_\infty$ défini sur Θ_∞^2, proche du noyau $\mathcal{K}_T$, sur lequel nous pouvons établir aisément des propriétés de bornitudes et de concavité locale sur la diagonale. Ces propriétés sont ensuite étendues à $\mathcal{K}_T$ par des conditions de proximité entre les deux noyaux. Typiquement, on considère le cadre $\lim_{T\to+\infty} \mathcal{K}_T = \mathcal{K}_\infty$.

Exemple 1.9 (Processus à temps discret sur une grille régulière). Poursuivons l'exemple 1.2. La suite de mesures $(\lambda_T, T \geq 2)$ converge alors par rapport à la topologie vague vers la mesure de Lebesgue Leb sur $\Theta_\infty = \mathbb{R}$. L'espace de Hilbert $H_\infty = L^2(\lambda_\infty)$ doté de son produit scalaire usuel apparaît ainsi comme l'espace limite. Considérons le modèle de pics gaussiens continûment paramétrés par leurs positions avec un paramètre d'échelle fixe $\sigma_0 > 0$. Le dictionnaire est donné par:

$$\Big(\varphi(\theta) = h\Big(\frac{\cdot - \theta}{\sigma_0}\Big),\, \theta \in \Theta\Big) \quad \text{avec} \quad h(t) = \mathrm{e}^{-t^2/2} \quad \text{et} \quad \Theta = \mathbb{R}.$$

On considère dans ce cas le noyau limite donné sur $\Theta_\infty^2 = \mathbb{R}^2$ par $\mathcal{K}_\infty(\theta,\theta') = h\Big(\frac{\theta-\theta'}{\sqrt{2}\,\sigma_0}\Big)$.

On note $|\Theta_T|_{\mathfrak{d}_T}$ la longueur de l'intervalle compact $\Theta_T \subset \mathbb{R}$ par rapport à la distance $\mathfrak{d}_T$ dérivée du noyau $\mathcal{K}_T$. La contribution principale du chapitre s'énonce comme suit. Nous renvoyons aux théorèmes 2.2.1 et 2.2.5 ainsi qu'aux propositions 2.7.4 et 2.7.5 pour des énoncés détaillés.

Théorème 1.9.1 (Version informelle)**.** *Supposons que l'on observe un élément aléatoire y de H_T issu du modèle (1.1) de paramètres inconnus $s \in \mathbb{N}^*$ avec $s \leq K$, $\beta^\star \in (\mathbb{R}^*)^s$ et $\vartheta^\star = (\theta_1^\star, \cdots, \theta_s^\star) \in \Theta_T^s$, où Θ_T est un intervalle compact de $\mathbb{R}$, tel que:*

(i) Le bruit est admissible au sens de l'hypothèse 1.1.1 avec $\sigma > 0$ et $\Delta_T > 0$.

(ii) La fonction φ_T générant le dictionnaire est suffisamment régulière.

(iii) Le noyau approximant $\mathcal{K}_\infty$ est régulier et localement concave sur la diagonale.

(iv) Le noyau approximant $\mathcal{K}_\infty$ est proche de $\mathcal{K}_T$.

(v) Les paramètres non-linéaires sont suffisamment séparés de sorte que pour $\theta \neq \theta' \in \mathcal{Q}^\star = \{\theta_i^\star,\, 1 \leq i \leq s\}$:

$$\mathfrak{d}_T(\theta, \theta') > \Sigma_{s,T},$$

Alors, il existe des constantes positives $\mathcal{C}_i$ avec $i = 0, \cdots, 4$, dépendant uniquement du noyau $\mathcal{K}_\infty$ défini sur Θ_∞ et de r, telles que pour $\tau > 1$ et un paramètre de pénalisation:

$$\kappa \geq \mathcal{C}_1 \sigma \sqrt{\Delta_T \log \tau},$$

on ait une borne de prédiction associée aux estimateurs $\hat{\beta}$ et $\hat{\vartheta}$ définis par (1.13), donnée par:

$$\left\| \beta^\star \, \Phi_T(\vartheta^\star) - \hat{\beta} \, \Phi_T(\hat{\vartheta}) \right\|_T \leq \mathcal{C}_0 \sqrt{s}\, \kappa,$$

avec probabilité au moins

$$1 - \mathcal{C}_2 \left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau \sqrt{\log \tau}} \vee \frac{1}{\tau} \right).$$

De plus, avec la même probabilité, la différence des normes ℓ_1 de $\hat{\beta}$ et $\beta^\star$ est bornée par:

$$\left| \| \hat{\beta} \|_{\ell_1} - \| \beta^\star \|_{\ell_1} \right| \leq \mathcal{C}_3 \, \kappa \, s.$$

On a également avec les ensembles d'indices $S(r)$ et $S_k(r)$ définis par (1.5):

$$\sum_{k=1}^{s} \left| |\beta_k^\star| - \sum_{\ell \in S_k(r)} |\hat{\beta}_\ell| \right| \quad + \quad \sum_{k=1}^{s} \left| \beta_k^\star - \sum_{\ell \in S_k(r)} \hat{\beta}_\ell \right| \quad + \quad \left\| \hat{\beta}_{S(r)^c} \right\|_{\ell_1} \leq \mathcal{C}_4 \, \kappa \, s. \tag{1.17}$$

Remarque 1.10. Une borne explicite en les paramètres du problème est donnée dans la section 2.8.2 du chapitre 2 sur la quantité $\Sigma_{s,T}$ pour le modèle de déconvolution de pics gaussiens.

Notons que les contrôles donnés dans le théorème 1.9.1 ne dépendent pas de la borne K sur la parcimonie s. La borne de prédiction ainsi que les deux dernières inégalités de (1.17) étendent des résultats jusqu'à présent limités au cas spécifique d'un dictionnaire constitué d'exponentielles complexes continûment paramétrées par leurs fréquences (voir [Boyer et al., 2017, Tang et al., 2015]). Plutôt que d'utiliser des contrôles sur $\|w_T\|_T$ comme dans les travaux précurseurs de [Duval and Peyré, 2015, Poon et al., 2021], le résultat s'appuie sur des contrôles de queues de distribution de suprema de processus gaussiens de la forme $\sup_{\Theta_T} \langle f(\theta), w_T \rangle_T$ pour certaines fonctions f, définies sur Θ_T à valeurs dans H_T, construites à partir de la fonction φ_T et de ses deux premières dérivées (voir Section 2.11.1 du chapitre 2).

Dans la lignée de nombreux travaux conduits au sein des communautés de l'acquisition comprimée et de la la super-résolution ([Candès and Fernandez-Granda, 2014, Candès and

Fernandez-Granda, 2013] entre autres), nos bornes reposent sur l'existence de certificats présentés en détail dans la Section 2.6 du chapitre 2. De telles fonctions peuvent être construites à condition que les paramètres non linéaires du mélange soient suffisamment séparés relativement à la métrique riemannienne $\mathfrak{d}_T$ associée au noyau $\mathcal{K}_T$ (voir Point (v) du théorème 1.9.1). Nous les construisons explicitement dans les propositions 2.7.4 et 2.7.5 dans l'esprit de [Poon et al., 2021].

La borne sur le risque de prédiction obtenue dans le théorème 1.9.1 correspond (à un facteur logarithmique près) à celle atteinte par l'estimateur Lasso dans le cas linéaire, c'est-à-dire lorsque $\vartheta^\star$ est connu et n'a pas besoin d'être estimé, voir la remarque 2.2.2.

1.10.1 Contributions du chapitre 3

Le contenu de ce chapitre est issu de [Butucea et al., 2022c].

Dans ce chapitre, nous étendons les résultats du chapitre 2 afin de couvrir le modèle (1.3) dans lequel plusieurs éléments de H_T sont observés. Rappelons que les observations sont indexées par un ensemble $\mathcal{Z}$ et peuvent être pondérées à l'aide d'une mesure finie positive ν sur $\mathcal{Z}$. Nous établissons une borne supérieure, valable avec une certaine probabilité, sur l'erreur de prédiction associée aux estimateurs solutions du problème (1.15) régularisé par une norme mixte $(\ell_1, L^p(\nu))$ avec $p \in [1,2]$. Le résultat principal de ce chapitre s'énonce comme suit (voir le théorème 3.3.1 pour un énoncé détaillé).

Théorème 1.10.1 (Version informelle)**.** *Soit $T \in \mathbb{N}$. Soit $p \in [1,2]$ et $q \in [2,+\infty]$ tels que $1/p + 1/q = 1$. Lorsque $p = 1$, on suppose que $\mathcal{Z}$ est fini. Supposons que l'on observe un élément Y de $L_T = L^2(\nu, H_T)$ issu du modèle (1.3) de paramètres inconnus $s \in \mathbb{N}^*$ tel que $s \leq K$, $B^\star \in L^2(\nu, \mathbb{R}^s)$ et $\vartheta^\star = (\theta_1^\star, \cdots, \theta_s^\star) \in \Theta_T^s$ et où le bruit W_T appartient à $L^q(\nu, H_T)$. Supposons que les points (ii)-(v) du théorème 1.9.1 soient vérifiés.*

Alors, il existe des constantes finies positives $\mathcal{C}$, $\mathcal{C}_0$ (dépendant de r et du noyau $\mathcal{K}_\infty$ défini sur Θ_∞) telles que l'erreur de prédiction associée aux estimateurs $\hat{B}$ et $\hat{\vartheta}$ définis par (1.15) pour un paramètre de pénalisation $\kappa > 0$, est donnée par:

$$\frac{1}{\sqrt{\nu(\mathcal{Z})}} \left\| B^\star \, \Phi_T(\vartheta^\star) - \hat{B} \, \Phi_T(\hat{\vartheta}) \right\|_{L_T} \leq \mathcal{C}_0 \sqrt{s} \, \nu(\mathcal{Z})^{\frac{1}{p}} \, \kappa,$$

avec probabilité au moins

$$1 - \sum_{i=0}^{2} \mathbb{P}\left(M_i > \mathcal{C} \, \kappa \, \nu(\mathcal{Z}) \right),$$

où M_i est défini par:

$$M_i = \sup_{\theta \in \Theta_T} \left\| \left\langle W_T(\cdot), \phi_T^{[i]}(\theta) \right\rangle_T \right\|_{L^q(\nu)}, \quad \textit{pour} \quad i = 0, 1, 2,$$

où la dérivée covariante $\phi_T^{[i]} = \tilde{D}_{i,\mathcal{K}_T}[\phi_T]$ est donnée dans la section 2.4 du chapitre 2.

Le résultat ci-dessus repose sur l'existence de certificats généralisant ceux introduits dans le chapitre 2 (voir la section 3.4 pour une présentation complète de ces objets). Sous des conditions identiques à celles du théorème 1.9.1 (dont une condition de séparation sur les paramètres non linéaires $\mathcal{Q}^\star$), nous montrons dans les propositions 3.4.1 et 3.4.2 que l'existence de certificats est vérifiée.

En pratique, il n'existe pas de manière simple de contrôler les queues de distribution des variables aléatoires M_i. Néanmoins, lorsque l'ensemble $\mathcal{Z}$ est fini et que ν est la mesure de comptage sur $\mathcal{Z}$, des contrôles peuvent être donnés pour $p = 1$ et $p = 2$. Pour $p = 1$, il s'agit de contrôler des queues de suprema de processus gaussiens réguliers tandis que dans le cas $p = 2$, cela revient à contrôler la queue du supremum d'un processus du χ^2 (voir la section 3.9.2 du chapitre 3). Il est ensuite possible d'en déduire des contrôles sur les queues de distribution

des variables aléatoires M_i associées à des paires conjuguées (p, q) dans $[1, 2] \times [2, \infty]$ à l'aide d'inégalités d'interpolation.

Nous donnons dans le corollaire suivant une borne sur le risque de prédiction analogue à celle minimax établie pour les modèles de régression linéaire multi-tâches. Nous tirons ainsi profit de la structure commune des signaux afin d'améliorer leur reconstruction.

Corollaire 1.11 (**Cas $\mathcal{Z}$ fini et** $p = 2$ (version informelle)). *Soit $T \in \mathbb{N}$ et fixons $p = q = 2$. Supposons que* $\mathrm{Card}(\mathcal{Z}) = n < +\infty$ *et que la mesure ν soit la mesure de comptage sur $\mathcal{Z}$. Supposons que le bruit W_T soit un élément de $L^q(\nu, H_T)$ (identifié à $H_T^{\otimes n}$) tel que les éléments $W_T(z)$ de H_T, avec $z \in \mathcal{Z}$, soient indépendants, identiquement distribués et satisfassent individuellement l'hypothèse 1.1.1.*

Alors, en reprenant toutes les notations du théorème 1.10.1 et en supposant que ses hypothèses soient vérifiées, il existe des constantes finies positives $\mathcal{C}_0$, $\mathcal{C}_1$, $\mathcal{C}_2$, (dépendant de $\mathcal{K}_\infty$ et de r) telles que pour tout $\tau > 1$ et paramètre de pénalisation:

$$\kappa \geq \mathcal{C}_1 \sigma \sqrt{\frac{\Delta_T}{n}} \left(1 + \sqrt{1 + \frac{\log(\tau)}{n}}\right),$$

on ait la borne de prédiction suivante pour les estimateurs $\hat{B}$ and $\hat{\vartheta}$ définis par (1.15):

$$\frac{1}{\sqrt{n}} \left\| B^\star \, \Phi_T(\vartheta^\star) - \hat{B} \, \Phi_T(\hat{\vartheta}) \right\|_{L_T} \leq \mathcal{C}_0 \sqrt{s \, n} \, \kappa,$$

avec probabilité au moins $1 - \mathcal{C}_2 \left(\frac{1}{\tau} + \frac{|\Theta_T|_{\mathfrak{d}_T} F(n)}{\sqrt{\tau}}\right)$ où $(F(n), \, n \geq 1)$ est une suite convergeant vers 0 à une vitesse de l'ordre de $\sqrt{n}\, \mathrm{e}^{-n/2}$ lorsque le nombre n de signaux croît.

La borne sur le risque de prédiction obtenue dans le corollaire 1.11 correspond (à un facteur logarithmique près) à celle atteinte par l'estimateur group-Lasso lorsque les paramètres non linéaires sont connus et n'ont pas besoin d'être estimés, voir la remarque 3.3.5. On montre ainsi que lorsque les signaux ont une structure commune, la reconstruction simultanée via (1.15) est plus performante que la reconstruction individuelle.

1.11.1 Contributions du chapitre 4

Le contenu de ce chapitre est issu de [Butucea et al., 2022b].

Ce chapitre est motivé par l'étude du modèle de déconvolution de pics utile en spectroscopie infrarouge. Dans ce modèle, une combinaison linéaire de pics continûment paramétrés par leurs positions est observée avec un processus de bruit additif, supposé ici gaussien. Dans ce chapitre, l'espace d'observation est $H_T = L^2(\lambda_T)$ où $(\lambda_T, \, T \geq 1)$ est une suite de mesures σ-finies, discrètes ou continues, sur le tore ou sur $\mathbb{R}$, convergeant vers la mesure de Lebesgue. Nous considérons des dictionnaires issus d'un modèle de translation dépendant d'un paramètre d'échelle σ_T:

$$\Big(\varphi_T(\theta) = h(\theta - \cdot, \sigma_T), \, \theta \in \Theta\Big)$$

où h est une fonction à valeurs réelles définie sur $\Theta \times \mathfrak{S}$, régulière par rapport à sa première variable et normalisée de telle sorte que $\|h(\cdot, \sigma_T)\|_{L^2(\mathrm{Leb})} = 1$, et où σ_T est un élément de l'ensemble des paramètres d'échelle admissibles $\mathfrak{S}$. Ce type de dictionnaire permet notamment de traiter la reconstruction de sources ponctuelles sur le tore à partir de leur convolution avec un filtre passe-bas dont la fréquence de coupure f_c peut varier. Typiquement, cela couvre l'exemple, fréquemment utilisé en super-résolution, où le signal observé est une somme de mesures de Dirac sur le tore, convolée avec un noyau de Dirichlet.

Lorsque que le paramètre d'échelle σ_T varie avec T, il arrive que le noyau limite, donné par $\mathcal{K}_\infty := \lim_{T\to+\infty} \mathcal{K}_T$, soit dégénéré, c'est-à-dire nul presque partout. Plutôt que d'utiliser les noyaux limites, nous considérons des noyaux approximants de la forme:

$$\mathcal{K}_T^{\text{prox}} : (\theta, \theta') \mapsto F(|\theta - \theta'|/\sigma_T),$$

où F est une fonction régulière, paire, à valeurs réelles, définie sur $\mathbb{R}$ et où $|\theta - \theta'|$ désigne la distance euclidienne entre deux éléments θ et θ' du tore ou de $\mathbb{R}$. Le choix de la fonction F est déterminé de sorte que $\mathcal{K}_T$ et $\mathcal{K}_T^{\text{prox}}$ soient proches.

Exemple 1.12 (Processus à temps discret sur une grille régulière : pics gaussiens). Poursuivons l'exemple 1.2. Dans le cas des pics gaussiens où

$$h(t,\sigma) \mapsto \frac{\exp(-t^2/2\sigma^2)}{\pi^{1/4}\sigma^{1/2}} \quad \text{est définie sur} \quad \Theta \times \mathfrak{S} = \mathbb{R} \times \mathbb{R}_+^*,$$

on considère dans la section 4.5 du chapitre 4, la fonction F donnée par:

$$F(t) = \exp(-t^2/4).$$

Exemple 1.13 (Processus à temps continu sur le tore : le filtre passe-bas). Poursuivons l'exemple 1.3. Dans le cas du filtre passe-bas où

$$h(t,\sigma) = \frac{\sin(T\pi t)}{\sqrt{T}\,\sin(\pi t)} \quad \text{est définie pour} \quad t \in \Theta = \mathbb{R}/\mathbb{Z} \quad \text{et} \quad \sigma = \frac{1}{T}, \quad T \in 2\mathbb{N}^* + 1,$$

on considère dans la section 4.6 du chapitre 4, la fonction F donnée par:

$$F(t) = \frac{\sin(\pi t)}{\pi t} \quad \text{pour} \quad t \in \mathbb{R}.$$

Dans un premier temps nous étendons les résultats de prédiction du chapitre 2 en prenant en compte ici que le paramètre d'échelle peut varier. Les bornes de prédiction obtenues sont analogues à celles du théorème (1.9.1). Cependant, nous améliorons la distance minimale requise entre deux paramètres de position consécutifs. Définissons pour un sous ensemble A de $\mathbb{R}$ ou du tore, l'ensemble $A^s(\delta)$ des vecteurs de A^s dont les composantes sont séparées deux à deux d'une distance δ au sens de la métrique euclidienne, plus formellement:

$$A^s(\delta) = \Big\{(\theta_1, \cdots, \theta_s) \in A^s : \ |\theta_\ell - \theta_k| > \delta \quad \text{pour tous les indices distincts} \quad k, \ell \in \{1, \ldots, s\}\Big\}.$$

avec la convention que pour $s = 0, 1$: $A^0(\delta) = \{0\}$ et $A^1(\delta) = A$. La condition de séparation requise entre les paramètres de position $\vartheta^\star = (\theta_1^\star, \cdots, \theta_s^\star)$ est alors:

$$\vartheta^\star \in \Theta_T^s\left(\sigma_T\, \Sigma_s\right),$$

où Σ_s est une quantité dépendant *a priori* de la parcimonie s. Dans le cas général, pour une parcimonie s fixée, la séparation est de l'ordre du paramètre d'échelle σ_T. Nous verrons que pour certains modèles, la séparation est de l'ordre de σ_T quelque soit s (voir section 4.5).

Ensuite, nous proposons un test afin de décider si l'observation y est issue d'une combinaison linéaire donnée de pics. Le but est de déterminer si le signal inconnu $\beta^\star \Phi_T(\vartheta^\star)$ est égal à un signal de référence $\beta^0 \Phi_T(\vartheta^0)$ pour certains vecteurs connus $\beta^0 \in (\mathbb{R}^*)^{s^0}$ et $\vartheta^0 \in \Theta_T^{s^0}(\delta_0)$ où $\delta^0 \in \mathbb{R}_+$. Cela revient à tester les hypothèses:

$$\begin{cases} H_0 : & (\beta^\star, \vartheta^\star) \in (\mathbb{R}^*)^s \times \Theta_T^s(\delta^\star) \quad \text{tels que} \quad \beta^\star \Phi_T(\vartheta^\star) = \beta^0 \Phi_T(\vartheta^0), \\ H_1(\rho) : & (\beta^\star, \vartheta^\star) \in (\mathbb{R}^*)^s \times \Theta_T^s(\delta^\star) \quad \text{tels que} \quad \left\|\beta^\star \Phi_T(\vartheta^\star) - \beta^0 \Phi_T(\vartheta^0)\right\|_{L^2(\lambda_T)} \geq \rho. \end{cases} \tag{1.18}$$

Nous montrons que ce cadre inclut le problème de détection de signal formulé par les hypothèses (1.16). En outre, sous certaines conditions, détaillées plus bas dans le théorème 1.13.1, nous montrons que l'hypothèse nulle implique l'égalité des coefficients linéaires et des paramètres non-linéaires du signal observé et du signal de référence. Afin de distinguer les hypothèses H_0 et $H_1(\rho)$ introduites ci-dessus, nous construisons un test Ψ, *i.e.*, une fonction mesurable de l'observation y à valeurs dans $\{0,1\}$. Les erreurs de type I (rejeter H_0 à tort) et II (accepter H_0 à tort) sont respectivement:

$$\sup_{(\beta^\star,\vartheta^\star)\in H_0} \mathbb{E}_{(\beta^\star,\vartheta^\star)}[\Psi] \quad \text{et} \quad \sup_{(\beta^\star,\vartheta^\star)\in H_1(\rho)} \mathbb{E}_{(\beta^\star,\vartheta^\star)}[1-\Psi].$$

Le risque de test maximal pour Ψ est alors la somme des quantités précédentes, à savoir:

$$R_\rho(\Psi) = \sup_{(\beta^\star,\vartheta^\star)\in H_0} \mathbb{E}_{(\beta^\star,\vartheta^\star)}[\Psi] + \sup_{(\beta^\star,\vartheta^\star)\in H_1(\rho)} \mathbb{E}_{(\beta^\star,\vartheta^\star)}[1-\Psi],$$

et le risque de test minimax est:

$$R_\rho^\star = \inf_\Psi R_\rho(\Psi),$$

où l'infinimum porte sur toutes les fonctions mesurables de $L^2(\lambda_T)$ vers $\{0,1\}$. La séparation minimax du problème de test est définie pour $\alpha \in (0,1)$ comme la plus petite séparation possible permettant de tester l'hypothèse nulle et son alternative avec un risque minimax inférieur à un niveau α:

$$\rho^\star(\alpha) = \inf\{\rho > 0 : R_\rho^\star \le \alpha\}.$$

On donne dans ce chapitre une borne supérieure non asymptotique sur la séparation minimax permettant de distinguer deux signaux distincts. Le résultat suivant regroupe les corollaires 4.3.2 et 4.3.5 et le lemme 4.2.4.

Théorème 1.13.1 (Version informelle)**.** *Soient $T \in \mathbb{N}$, $s^0 \in \mathbb{N}$ et $K \in \mathbb{N}$. Considérons l'élément aléatoire y de $L^2(\lambda_T)$ issu du modèle (1.1) de paramètres inconnus $s \in \mathbb{N}$ tel que $s \le K$, $\beta^\star \in (\mathbb{R}^*)^s$ et $\vartheta^\star \in \Theta_T^s(\delta^\star)$ avec $\delta^\star \ge \sigma_T \Sigma_s$. Soient $\beta^0 \in (\mathbb{R}^*)^{s^0}$ et $\vartheta^0 \in \Theta_T^{s^0}(\delta^0)$ avec $\delta_0 \ge \sigma_T \Sigma_{s^0}$. Supposons que:*

(i) Le bruit est admissible au sens de l'hypothèse 1.1.1 avec $\overline{\sigma} > 0$ et $\Delta_T > 0$. On a aussi $\Xi_T := \mathrm{Var}\left(\|w_T\|^2_{L^2(\lambda_T)}\right) < +\infty$.

(ii) La fonction φ_T générant le dictionnaire est suffisamment régulière.

(iii) La fonction F est régulière et localement concave en 0.

(iv) Le noyau approximant $\mathcal{K}_T^{prox}$ est proche de $\mathcal{K}_T$.

Alors, à condition que $|\Theta_T|/\sigma_T \ge 1$, il existe des constantes finies positives c et C, (dépendant de r et de la fonction F) telles que la séparation minimax pour le problème de test (1.18) vérifie pour $\alpha \in (0,1)$:

$$\rho^\star(\alpha) \le C \min\left(\left(\frac{\Xi_T}{\alpha}\right)^{1/4}, \overline{\sigma}\sqrt{(s \vee s^0 \vee 1)\Delta_T \log\left(\frac{2\,c\,|\Theta_T|}{\alpha\,\sigma_T}\right)}\right),$$

De plus, sous l'hypothèse nulle H_0, on a, à une permutation identique près sur les composantes de $\beta^\star$ et $\vartheta^\star$:

$$s = s^0, \quad \beta^\star = \beta^0 \quad et \quad \vartheta^\star = \vartheta^0.$$

La borne supérieure obtenue sur la séparation minimax donne lieu à deux régimes selon que le signal observé et le signal de référence sont parcimonieux ou non. La procédure de test permettant d'atteindre cette borne supérieure est donnée explicitement. Dans le cadre de la détection de signaux où l'hypothèse nulle est $\beta^\star \equiv 0$, nous déduisons ainsi des bornes

supérieures sur l'énergie minimale qu'un signal doit avoir pour être détecté en présence de bruit. Il s'avère que, dans cette configuration, notre borne supérieure correspond (à un facteur logarithmique près) à une borne inférieure asymptotique établie dans [Ingster et al., 2010] pour le modèle linéaire de grande dimension associé à un dictionnaire fini.

Nous testons également si les composantes actives du signal observé sont incluses dans une collection finie et connue de composantes et si les signes avec lesquels elles apparaissent sont ceux prescrits. Plus formellement, soient $\mathcal{Q}^\star = \{\theta_1^\star, \cdots, \theta_s^\star\}$ l'ensemble des paramètres inconnus du signal et $\mathcal{Q}^0 = \{\theta_1^0, \cdots, \theta_{s^0}^0\}$ un sous ensemble connu de Θ_T de cardinal s^0. Nous voulons décider si pour chaque $\epsilon = \pm 1$, l'ensemble inconnu $\mathcal{Q}^{\star,\epsilon} = \{\theta_k^\star \in \mathcal{Q}^\star : \epsilon\,\beta_k^\star > 0\}$ est un sous-ensemble de $\mathcal{Q}^{0,\epsilon} = \{\theta_k^0 \in \mathcal{Q}^0 : \epsilon\, v_k^0 > 0\}$, où $v^0 \in \{-1,1\}^{s^0}$ est un vecteur de signe. Lorsque $s_0 = 0$ notons que nous sommes ramenés à la détection de signal déjà traitée. Ainsi nous supposerons que $s_0 \geq 1$. Lorsque l'ensemble $\mathcal{Q}^{0,-}$ est vide, le test consiste à décider si les composantes du signal sont incluses dans $\mathcal{Q}^{0,+}$ et sont bien toutes associées à un coefficient linéaire positif. Pour séparer l'hypothèse nulle H_0 de son alternative H_1, nous introduisons une mesure de discrépance s'annulant uniquement lorsque les paramètres $(\beta^\star, \vartheta^\star)$ appartiennent à H_0. Cette mesure de discrépance est définie pour un paramètre $r > 0$ par:

$$\mathcal{D}_{T,r}(\beta^\star, \vartheta^\star, v^0, \vartheta^0) = \sum_{\epsilon\in\{-1,+1\}} \sum_{\substack{k \quad t.q \\ \epsilon\beta_k^\star > 0, \\ \mathfrak{d}_T(\theta_k^\star, \mathcal{Q}^{0,\epsilon}) \leq r}} |\beta_k^\star|\, \mathfrak{d}_T(\theta_k^\star, \mathcal{Q}^{0,\epsilon})^2 + \sum_{\epsilon\in\{-1,+1\}} \sum_{\substack{k \quad t.q \\ \epsilon\beta_k^\star > 0, \\ \mathfrak{d}_T(\theta_k^\star, \mathcal{Q}^{0,\epsilon}) > r}} |\beta_k^\star|.$$

Nous donnons une borne supérieure sur la séparation minimax $\rho^\star$ liée au test. La statistique de test introduite et étudiée dans ce contexte fait explicitement appel aux certificats donnés dans [Butucea et al., 2022a] pour établir des bornes sur les risques de prédiction des estimateurs de $(\beta^\star, \vartheta^\star)$.

1.13.1 Contributions du chapitre 5

Le contenu de ce chapitre est issu de [Butucea et al., 2021].

La spectroscopie infrarouge vise à mesurer l'interaction de rayonnements infrarouges avec la matière. Les spectres obtenus par cette méthode renseignent sur la présence de composés chimiques ou de groupes fonctionnels dans un échantillon. Dans l'industrie, et notamment à Électricité de France (EDF), ces informations sur la composition de la matière sont essentielles pour prévenir les défaillances de matériaux.

Lorsqu'un grand nombre de spectres doit être analysé, une procédure automatique est nécessaire. Une nouvelle approche est proposée dans ce chapitre pour analyser automatiquement et simultanément un ensemble de spectres infrarouges. Les spectres considérés présentent de nombreux pics résultant de l'absorption d'un rayonnement infrarouge par un des composés chimiques du matériau. Dans la lignée de travaux menés en spectroscopie ([Aragoni et al., 1995], [Hollas, 2004]), nous modélisons les spectres par des combinaisons linéaires de pics dont la dispersion et la position sont continûment paramétrées. Des profils gaussiens ou de Lorentz peuvent être considérés pour les pics. Plus les amplitudes de ces derniers sont grandes, plus les composés chimiques associés sont concentrés. Lorsqu'on cherche à détecter des défauts dans une infrastructure, il est courant de disposer d'un grand nombre d'échantillons d'un même matériau. Ceux-ci partagent alors de nombreux composés chimiques. En conséquence, les spectres obtenus présentent une structure commune: les positions et les dispersions des pics sont partagées par l'ensemble des spectres tandis que leurs amplitudes sont propres à chacun. Afin de récupérer les paramètres des pics et les amplitudes associées, nous résolvons le problème d'optimisation pénalisé (1.15). Nous utilisons, dans la lignée des travaux de [Denoyelle et al., 2020], [Boyd et al., 2017] et [Golbabaee and Poon, 2022], un algorithme alternant

des étapes convexes (pour estimer les amplitudes des pics) et des étapes non convexes (pour estimer les positions et les dispersions des pics). Une bibliothèque Python implémentant cet algorithme a été développée et mise à disposition des chercheurs d'EDF.

Nous étudions les performances numériques de l'algorithme sur des spectres infrarouges de revêtements en polychloroprène vieillis en milieu marin. Les emplacements des pics trouvés par l'algorithme sont cohérents avec ceux établis par des travaux antérieurs dans le domaine de la chimie (voir [Tchalla et al., 2017]). Notre méthode permet d'identifier de manière automatique les composés chimiques impliqués dans le processus de vieillissement du matériau et de regrouper les échantillons de polychloroprène selon leurs niveaux d'usure.

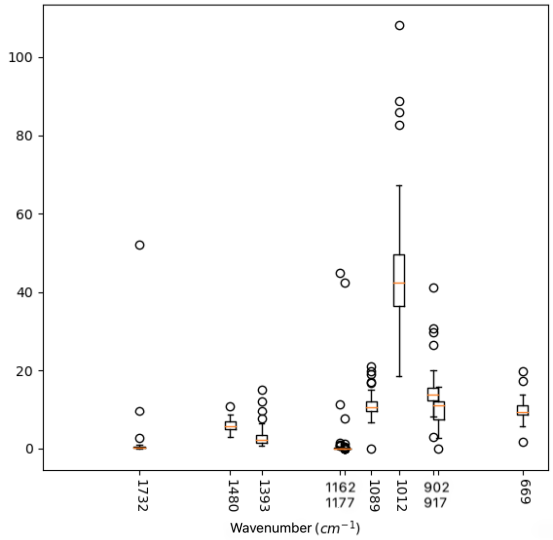

Figure 1.2 – Dispersion des amplitudes associées aux 10 principaux pics d'absorption (en termes de la norme euclidienne des amplitudes estimées parmi tous les spectres) déterminés par notre algorithme pour les spectres de polychloroprène vieilli en milieu marin présentés dans la figure 1.1.

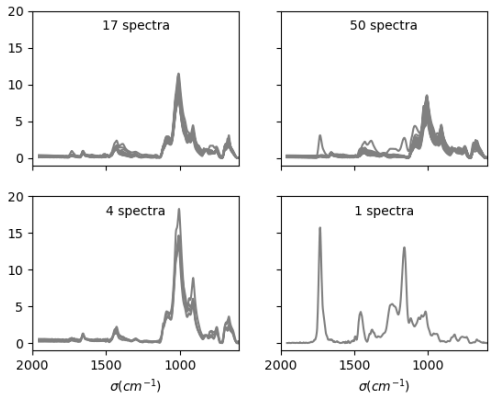

Figure 1.3 – Regroupement des spectres de polychloroprène vieilli en milieu marin présentés dans la figure 1.1 par niveaux d'usure.

2 Off-the-grid learning of sparse mixtures from a continuous dictionary

Contents

Preamble

We consider a general non-linear model where the signal is a finite mixture of an unknown, possibly increasing, number of features issued from a continuous dictionary parameterized by a real non-linear parameter. The signal is observed with Gaussian (possibly correlated) noise in either a continuous or a discrete setup. We propose an off-the-grid optimization method, that is, a method which does not use any discretization scheme on the parameter space, to estimate both the non-linear parameters of the features and the linear parameters of the mixture.

We use recent results on the geometry of off-the-grid methods to give minimal separation on the true underlying non-linear parameters such that interpolating certificate functions can be constructed. Using also tail bounds for suprema of Gaussian processes we bound the prediction error with high probability. Assuming that the certificate functions can be constructed, our prediction error bound is up to log $-$factors similar to the rates attained by the Lasso predictor in the linear regression model. We also establish convergence rates that quantify with high probability the quality of estimation for both the linear and the non-linear parameters.

The material of this chapter has been released in [Butucea et al., 2022a].

2.1 Introduction

2.1.1 Model and method

Assume we observe a random element y of an Hilbert space and we consider a signal-plus-noise structure for the observation y, where the noise is distributed according to a centered Gaussian process. The signal is modeled as a mixture model, by a linear combination of at most K features of the form $\varphi(\theta)$ for some parameters $\theta \in \Theta$, where $\Theta \subseteq \mathbb{R}$ is an interval of parameters and φ is a smooth function defined on Θ and taking values in the Hilbert space. We denote by $(\varphi(\theta), \theta \in \Theta)$ the continuous dictionary.

In order to capture a great variety of examples, we shall assume there exists a Hilbert space H_T, endowed with the scalar product $\langle \cdot, \cdot \rangle_T$ and the norm $\|\cdot\|_T$, where T is a parameter belonging to $\mathbb{N}$, such that: the observed process y belongs to H_T; for all $\theta \in \Theta$, the feature $\varphi_T(\theta)$ (which may depend on T) belongs to H_T and is non degenerate, *i.e.* $\|\varphi_T(\theta)\|_T$ is finite and non zero; the noise process w_T, which might also depend on the parameter T is a centered Gaussian process belonging to H_T. In the next example, the parameter T is understood as an amount of information, and, for T large, the Hilbert space H_T can be seen as an approximation of a limit Hilbert space.

Example 2.1.1 (Observations on a regular grid). Consider a real-valued process y observed over a regular grid $t_1 < \cdots < t_T$ on $[0,1]$, with $t_j = j/T$ and $T \in \mathbb{N}^*$, and the noise given by centered Gaussian random variables, say $G_1, \ldots, G_T$. Assuming that all the observations have the same weight amounts to considering y as an element of the Hilbert space $H_T = L^2(\lambda_T)$ of real valued functions defined on $[0,1]$ and square integrable with respect to the uniform probability measure λ_T on $\{t_1, \ldots, t_T\}$: $\lambda_T = T^{-1} \sum_{j=1}^T \delta_{t_j}$, where δ_x denotes the Dirac mass at x. In this formalism, the noise $w_T \in H_T$ is given by $w_T(t) = \sum_{j=1}^T G_j \mathbf{1}_{\{t_j\}}(t)$, where $\mathbf{1}_A$ denotes the indicator function of an arbitrary set A. Now, for T large, one can approximate the measure λ_T by the Lebesgue measure on $[0,1]$, say Leb. In various examples, it is also easier to compute the norms of the features and of their derivatives in the Hilbert space $L^2(\text{Leb})$. This amounts to seeing H_T as approximating Hilbert spaces of the fixed Hilbert space $L^2(\text{Leb})$.

Let us define the normalized function ϕ_T defined on Θ by:

$$\phi_T(\theta) = \frac{\varphi_T(\theta)}{\|\varphi_T(\theta)\|_T} \tag{2.1}$$

as well as the multivariate function Φ_T defined on Θ^K by:

$$\Phi_T(\vartheta) = (\phi_T(\theta_1), \ldots, \phi_T(\theta_K))^\top \quad \text{for} \quad \vartheta = (\theta_1, \cdots, \theta_K) \in \Theta^K.$$

We consider the model with unknown parameters $\beta^\star$ in $\mathbb{R}^K$ and $\vartheta^\star$ in Θ^K:

$$y = \beta^\star \Phi_T(\vartheta^\star) + w_T \quad \text{in } H_T. \tag{2.2}$$

We assume from now on that the unknown K dimensional vector $\beta^\star$ is sparse, *i.e* it has s non zero entries or, equivalently, $\beta^\star \in \mathcal{B}_0(s) = \left\{\beta \in \mathbb{R}^K,\ \|\beta\|_{\ell_0} = s\right\}$, where $\|\beta\|_{\ell_0}$ counts the number of non zero entries of the vector β. Let $S^\star$ be the support of $\beta^\star$:

$$S^\star = \text{Supp}(\beta^\star) = \{k \in \{1, \cdots, K\},\ \beta^\star_k \neq 0\},$$

and call $s = \text{Card}\, S^\star$ the sparsity parameter. We are interested in predicting observations and in recovering the unknown parameters. Let us denote in general by u_S the vector u in $\mathbb{R}^K$ restricted to the coordinates in S for any non-empty set $S \subseteq \{1, ..., K\}$. We estimate both the vector $\beta^\star_{S^\star}$ with unknown sparsity s and the vector $\vartheta^\star_{S^\star}$ with entries in some compact set Θ_T and containing the parameters of those functions from our continuous dictionary that appear

in the mixture model. Note that when applying the same permutation on the coordinates of $\beta^\star$ and the coordinates of $\vartheta^\star$, we obtain the same model. Thus, the vectors $\beta^\star$ and $\vartheta^\star$ are defined up to such a joint permutation. Moreover, we have $\beta^\star \Phi_T(\vartheta^\star) = \beta^\star_{S^\star} \Phi_T(\vartheta^\star)_{S^\star}$, where, by definition, $\Phi_T(\vartheta^\star)_{S^\star} = \Phi_T(\vartheta^\star_{S^\star})$. Our model is linear and sparse in $\beta^\star$ but it is non-linear in $\vartheta^\star$.

We make the following assumption on the noise process w_T, where the decay rate $\Delta_T > 0$ controls the noise variance decay as the parameter T grows and $\sigma > 0$ is the intrinsic noise level.

Assumption 2.1.1 (Admissible noise)**.** *Let $T \in \mathbb{N}$. The noise process w_T belongs to H_T a.s., and there exist a noise level $\sigma > 0$ and a decay rate $\Delta_T > 0$ such that for all $f \in H_T$, the random variable $\langle f, w_T \rangle_T$ is a centered Gaussian random variable satisfying:*

$$\mathrm{Var}\left(\langle f, w_T \rangle_T\right) \leq \sigma^2 \, \Delta_T \, \|f\|_T^2.$$

In our model, the parameter T may be understood as the amount of information that we have on the mixture, see Section 2.1.2 below. In the discrete case, see in particular Example 2.1.1, the amount of information grows as the frequency of the design points over which the process is observed increases; in the continuous case, it grows as the decay rate Δ_T of the noise variance decreases.

In order to recover the sparse vector $\beta^\star$ as well as the associated parameters $\vartheta^\star_{S^\star}$ (up to a permutation), we solve the following regularized optimization problem with a real tuning parameter $\kappa > 0$:

$$(\hat{\beta}, \hat{\vartheta}) \in \underset{\beta \in \mathbb{R}^K, \vartheta \in \Theta_T^K}{\operatorname{argmin}} \quad \frac{1}{2} \|y - \beta \Phi_T(\vartheta)\|_T^2 + \kappa \|\beta\|_{\ell_1}, \tag{2.3}$$

where the function Φ_T is assumed to be continuous and the set Θ_T on which the optimization of the non-linear parameters is performed is required to be a compact interval. Therefore the existence of at least a solution is guaranteed. The functional that we minimize in this problem is composed of a data fidelity term and a penalty term. The penalty is expressed with a ℓ_1-norm on the vector $\beta = (\beta_1, \ldots, \beta_K)$, *i.e* the sum of the absolute values of its coordinates: $\|\beta\|_{\ell_1} = \sum_{i=1}^K |\beta_i|$. This penalization is similar to that of the Lasso problem (also referred to as Basis pursuit) introduced in [Tibshirani, 1996] and extensively studied since then (see [Bühlmann and van de Geer, 2011] for a comprehensive survey). The optimization of the non-linear parameters is not performed on the whole set of parameters Θ but rather on a compact subset Θ_T indexed by the parameter T. Indeed, it may be necessary to restrict the set of parameters, *e.g.* in a finite mixture model where we consider a location parameter we can only recover those parameters within the support of the observations.

In the more general Beurling Lasso (BLasso) framework, one can rewrite the problem (2.3) in a measure setting. The actual solution $(\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_K), \hat{\vartheta} = (\hat{\theta}_1, \ldots, \hat{\theta}_K))$ of (2.3) is then seen as the atomic measure $\hat{\mu} = \sum_{i=1}^K \hat{\beta}_i \, \delta_{\hat{\theta}_i}$, where the amplitudes and the locations of the Dirac masses correspond respectively to the linear coefficients in the mixture and the parameters of the features. The measure $\hat{\mu}$ is also a solution of the BLasso problem when the latter admits atomic solutions composed of less than K atoms. This is in particular the case in the model presented in Section 2.1.2.1 where $H_T = \mathbb{R}^T$ and $K \geq T$ according to [Boyer et al., 2019]. However, to our knowledge, there are no such results when H_T is a general Hilbert space.

2.1.2 Examples of admissible noises

We give here various examples of discrete or continuous noise processes that satisfy our assumptions. They are frequently used in discrete regression models or continuous models like the Gaussian white noise model, see [Tsybakov, 2009] or [Giné and Nickl, 2016].

2.1.2.1 Discrete model with unweighted observations

Let $T \in \mathbb{N}^*$ and $H_T = \mathbb{R}^T$ endowed with the usual scalar product and the corresponding Euclidean norm $\|\cdot\|_T = \|\cdot\|_{\ell_2}$. If the noise w_T is a vector of T independent centered Gaussian random variables with variance σ^2, then Assumption 2.1.1 holds with an equality and $\Delta_T = 1$:

$$\mathrm{Var}(\langle f, w_T\rangle_T) = \sigma^2 \|f\|_{\ell_2}^2.$$

If w_T is a centered Gaussian vector of dimension T with each coordinate with variance σ^2, then Assumption 2.1.1 holds with Δ_T the spectral radius of the correlation matrix:

$$\mathrm{Var}(\langle f, w_T\rangle_T) \leq \sigma^2 \Delta_T \|f\|_{\ell_2}^2.$$

2.1.2.2 Discrete model with weighted observations

Assume the data set comes from the observations of a process y on a grid $t_1 < \cdots < t_T$ of size $T \in \mathbb{N}^*$. In this case, it might be pertinent to use a more general formalism (this is motivated by the case T large as in Example 2.1.1). In order to take into account possible different weights on the grid (which is legitimated when the grid is not regular), one can consider an atomic measure λ_T on $\mathbb{R}$ given by a (non-negative) linear combination of the Dirac masses on the grid, and the Hilbert space $H_T = L^2(\lambda_T)$ of real valued functions defined in $\mathbb{R}$ and square integrable with respect to the measure λ_T. We can then consider the noise given by the function $w_T = \sum_{j=1}^T G_j \mathbf{1}_{\{t_j\}}$ defined on $\mathbb{R}$, where $(G_1, \ldots, G_T)$ is a centered Gaussian vector with independent entries and common variance σ^2. In the particular case $\lambda_T(\mathrm{d}t) = \Delta_T \sum_{j=1}^T \delta_{t_j}(\mathrm{d}t)$ for some $\Delta_T > 0$ (in Example 2.1.1, $\Delta_T = 1/T$ and λ_T is a discrete approximation of the Lebesgue measure on $[0,1]$), we get that Assumption 2.1.1 holds with an equality:

$$\mathrm{Var}(\langle f, w_T\rangle_T) = \sigma^2 \Delta_T \|f\|_T^2.$$

In relation to the first model of Section 2.1.2.1, notice that in the present case $\|f\|_T = \sqrt{\Delta_T}\|f\|_{\ell_2}$, where the right-hand side is understood as the ℓ_2-norm (or Euclidean norm) of the vector $(f(t_1), \ldots, f(t_T))$.

2.1.2.3 Continuous model with truncated white noise or colored noise

Consider the set $\mathcal{C} = \mathcal{C}([0,1], \mathbb{R})$ of $\mathbb{R}$-valued continuous functions defined on $[0,1]$, an orthonormal base $(\psi_k, k \in \mathbb{N})$ of $L^2 = L^2([0,1], \mathrm{Leb})$ of elements of $\mathcal{C}$, where Leb is the Lebesgue measure on $[0,1]$. We simply denote by $\langle \cdot, \cdot \rangle_{L^2}$ the corresponding scalar product. Let $p = (p_k, k \in \mathbb{N})$ be a sequence of non-negative real numbers and set $\mathrm{Supp}(p) = \{k \in \mathbb{N} : \ p_k > 0\}$ its support. Let H_T be the completion of the vector space generated by the base $(\psi_k, k \in \mathrm{Supp}(p))$ (which is also the completion of $\mathcal{C}$ if p is positive and bounded), with respect to the scalar product:

$$\langle f, g\rangle_T = \sum_{k\in\mathbb{N}} p_k \langle f, \psi_k\rangle_{L^2} \langle g, \psi_k\rangle_{L^2}.$$

Notice that the Hilbert space H_T does not depend on the parameter T unless p depends on T. Let us recall that if $p \equiv 1$, that is, the sequence p is constant equal to 1, then $H_T = L^2$.

Let $\xi = (\xi_k, k \in \mathbb{N})$ be a weight sequence of non-negative real numbers such that the sequence $p\,\xi = (p_k\,\xi_k, k \in \mathbb{N})$ is summable. Consider the noise $w_T = \sum_{k\in\mathrm{Supp}\,(p)} \sqrt{\xi_k}\, G_k\, \psi_k$, where $(G_k, k \in \mathbb{N})$ are independent centered Gaussian random variables with variance σ^2. Notice Assumption 2.1.1 holds as $\|w_T\|_T^2 = \sum_{k\in\mathbb{N}} p_k\, \xi_k\, G_k^2$ is a.s. finite and, with $\Delta_T = \sup_{\mathbb{N}} p\,\xi$:

$$\mathrm{Var}(\langle f, w_T\rangle_T) = \sigma^2 \sum_{k\in\mathbb{N}} p_k^2\, \xi_k \langle f, \psi_k\rangle_{L^2}^2 \leq \sigma^2\, \Delta_T \|f\|_T^2.$$

Notice that the noise w_T does not depend on the parameter T unless p or ξ depends on T.

The truncated white noise model corresponds to $p \equiv 1$ and $\xi = (\xi_k = \mathbf{1}_{\{k \leq T\}}, k \in \mathbb{N})$. In this case $\Delta_T = 1$ and $\|w_T\|_T^2$ is a.s. of order $\sigma^2 T$ by the strong law of large numbers. The white noise corresponds to the limit case $T = +\infty$, which does not satisfy the hypothesis as a.s. its L^2-norm is infinite. Let us mention that the bounds given in the main theorems in Section 2.2 rely on $\|w_T\|_T$ being finite and not on its value.

Consider again $p \equiv 1$. Thanks to the Karhunen-Loève's decomposition, the scaled Brownian motion $w_T = C_T B$, with B the Brownian motion on $[0,1]$ and C_T a positive constant, corresponds to the base functions $\psi_k(t) = \sqrt{2}\sin((2k+1)\pi t/2)$ for $t \in [0,1]$ and the weights $\xi_k = 4C_T^2/(2k+1)^2\pi^2$ for $k \in \mathbb{N}$, and $\sigma^2 = 1$. In this case, we have $\langle f, w_T \rangle_T = C_T \int_0^1 f(s)B(s)\,\mathrm{d}s$ for $f \in L^2$ and Assumption 2.1.1 holds with $\sigma^2 = 1$ and $\Delta_T = \sup_{\mathbb{N}} p\,\xi = 4C_T^2/\pi^2$.

2.1.3 Previous work

The model (2.2) in the particular case where $\vartheta^\star$ is supposed given and the observations depend linearly on a vector $\beta^\star$ has long been studied in the literature. Assume for simplicity that $H_T = \mathbb{R}^T$ is the T-dimensional Euclidean space, so that $\Phi_T \in \mathbb{R}^{K \times T}$ is a matrix whose entries are known and can be either random or deterministic, $y \in \mathbb{R}^T$ is an observed vector and $w_T \in \mathbb{R}^T$ is a vector of noise (often assumed Gaussian). Even when K is larger than T the estimation of $\beta^\star$ is still consistent provided the vector $\beta^\star$ is sparse and a null space property is verified by the matrix Φ_T, or some sufficient condition saying that the lines of Φ_T are not too colinear (see [van de Geer, 2016] for a complete overview). The Lasso estimator [Tibshirani, 1996] or the Dantzig selector [Candès and Tao, 2007] are efficient to perfom such estimation and the quality of the estimation with respect to the dimension of the problem is now well known. The authors of [Bickel et al., 2009] have given bounds for the prediction error for both estimators.

We consider here a highly non-linear extension of this model that consists in assuming that the matrix $\Phi_T = \Phi_T(\vartheta^\star)$ depends non-linearly on a parameter $\vartheta^\star$ to be estimated. In our model (2.2), Φ_T is composed of K row vectors belonging to a parametric family or by K features belonging to a continuous dictionary and the observed data y may be either a vector or a function. This model has proven to be relevant in many fields such as microscopy, astronomy, spectroscopy, imaging or signal processing.

When the observation y belongs to a finite-dimensional Hilbert space and the dimension K is fixed and small compared to T, the model received attention several decades ago and gave rise to separable least square problems and resolution methods such as variable projection (see [Kaufman, 1975, Golub and Pereyra, 1973]). These papers mainly provided numerical methods but let us mention the consistency result in [Kneip and Gasser, 1988] for non-linear regression models.

On the contrary, when K is arbitrarily large many problems remain open. One of the natural ideas to estimate the underlying parameters could be to discretize the parameter space Θ and return to the study of a linear model. It would amount to considering a finite subfamily of $(\varphi(\theta),\, \theta \in \Theta)$ as in [Tang et al., 2013a] and deal with overcomplete dictionary learning techniques (also referred to as sparse coding, see [Olshausen and Field, 1997, Donoho et al., 2006]). In this case, sparse estimators for linear models such as the Lasso are available. However, in sparse spike deconvolution where the family $(\varphi(\theta), \theta \in \Theta)$ is a family of spikes parametrized by a location parameter, the authors of [Duval and Peyré, 2017a] have shown that in the presence of noise discretizing the space of parameters and solving a Lasso problem tends to produce clusters of spikes around the spikes one seeks to locate. That is why it is preferable to use off-the-grid methods. By off-the-grid, we mean that the methods employed do not use discretization schemes on the parameter set Θ. In [Duval and Peyré, 2015], the authors show that in presence of a small noise, the BLasso only induces a slight perturbation of the spikes locations and amplitudes and does not produce clusters. The

BLasso was introduced in [de Castro and Gamboa, 2012] and has been studied in many papers since then mostly by the compressed sensing and super-resolution communities ([Candès and Fernandez-Granda, 2013], [Azaïs et al., 2015] among many others). It is basically an off-the-grid extension of the classical Lasso for continuous dictionary learning. The optimization problem is formulated as a convex minimization over the space of Radon measures. In the BLasso framework, the dimension K in (2.2) is infinite and the linear coefficients and non-linear parameters are encoded by an atomic measure made of weighted Dirac functions. By solving a minimization problem over Radon measures, the aim is to recover an atomic measure. It raises the question of whether such a solution exists. In [Boyer et al., 2019] the question is answered by the affirmative when the observed data y belongs to a finite-dimensional Hilbert space H_T. When this is not the case, i.e. H_T is infinite dimensional, the question is open. In this chapter, we avoid the problem by assuming a bound K on the number of functions in the mixture and restricting the space over which the BLasso is perfomed to the atomic measures with at most K atoms. The numerical methods used to solve the BLasso such as the Sliding Frank-Wolfe algorithm ([Denoyelle et al., 2020] and [Butucea et al., 2021, Golbabaee and Poon, 2022] for applications in spectroscopy and imaging), also called the alternating descent conditional gradient method (see [Boyd et al., 2017]), and the conic particle gradient descent (see [Chizat, 2021]), seek a solution directly in the space of Dirac mixtures. Hence, our formulation (2.3) is closer to the way algorithms proceed. Let us mention that other methods such as Orthogonal Matching Pursuit (see [Elvira et al., 2021]) exist to tackle the problem of sparse learning from a continuous dictionary. Typically, the case of sparse spike deconvolution where the dictionary consists of Gaussian functions continuously parametrized by a location parameter is not included.

The study of the regression over a continuous dictionary in the framework of the BLasso has been quite specific to the dictionary considered. The literature first focused on the dictionary of complex exponential functions parametrized by their frequency $(\varphi(\theta) : t \mapsto \mathrm{e}^{i2\pi\langle t,\theta\rangle}, \theta \in \Theta)$ where Θ is the d-dimensional torus (see [Candès and Fernandez-Granda, 2014]). In [Boyer et al., 2017], a bound is given for the prediction error for this dictionary. The proof extends a previous result obtained in [Tang et al., 2015] for atomic norm denoising. What is particularly interesting is that the rates obtained for the prediction error almost reach the minimax rates achievable for linear models (see [Raskutti et al., 2011, Candès and Davenport, 2013]) provided that the frequencies are sufficiently separated. The separation condition between the non-linear parameters to estimate is inherent to the BLasso unless we assume the positivity of the linear parameters as in [Schiebinger et al., 2018].

For results on a wider range of dictionaries, let us highlight the work of [Duval and Peyré, 2015] that gives recovery and robustness to noise results for spike deconvolution. Let us also mention the recent work of [Bernstein et al., 2020] that generalizes some exact recovery results for a broader family of dictionaries as well as the paper [Bernstein and Fernandez-Granda, 2019] that gives robustness to noise guarantees for a family of shifted functions $(\varphi(\theta) = k(\cdot - \theta), \theta \in \Theta)$ of a given specific function k. In a density model that is a mixture of shifted functions, [De Castro et al., 2021] studies a modification of the BLasso by considering a weighted L^2 prediction error.

The case of non-translation invariant families remained for long intractable without very pessimistic separation conditions. In [Poon et al., 2021] the authors set a natural geometric framework to analyse the estimation problem. The separation condition between the parameters appears naturally in terms of a metric. In their paper, the design over which the observation are made is distributed according to a probability distribution. Their main result shows that in presence of noise the BLasso recovers a measure close to the one to be estimated with respect to a Wasserstein metric.

2.1.4 Contributions

This chapter adresses the problem of learning sparse mixtures from a continuous dictionary for a wide variety of regression models within a common framework. Indeed, we tackle a wide range of possible dictionaries of sufficiently smooth features, observation schemes and Gaussian noises with various structures. The observations are supposed to belong to a Hilbert space H_T. Continuous observations over an interval of $\mathbb{R}$ as well as discrete observations at given design points are therefore included in our framework. Furthermore, the Hilbert structure and the mild assumption we make on the noise, encompass a wide range of Gaussian noises. In particular, our framework allows to take into account the case of correlated Gaussian noise processes.

One of the main results of this chapter gives a high-probability bound for the prediction error:

$$\left\| \hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star \Phi_T(\vartheta^\star) \right\|_T,$$

where $(\hat{\beta}, \hat{\vartheta})$ is the solution of the optimization problem (2.3). Contrary to the BLasso optimization program over a set of measures whose result can be a diffuse measure, our formulation of the optimization problem has always a solution belonging to a finite set of values. Our prediction error bound matches (up to logarithmic factors and with high probability) that obtained in the linear case, that is when $\vartheta^\star$ is known and does not need to be estimated. We also give high-probability bounds on some loss functions comparing the estimators $\hat{\beta}$ and $\hat{\vartheta}$ given by (2.3) to the parameters $\beta^\star$ and $\vartheta^\star$, respectively. Our work extends results that were so far restricted to the specific case of a dictionary consisting of complex exponentials continuously parameterized by their frequencies (see [Boyer et al., 2017, Tang et al., 2015]). When the optimization problem produces a cluster of features to approximate an element of the mixture, we also show that there can be no compensation between the amplitudes of the features involved.

Following some work in compressed sensing and super-resolution ([Candès and Fernandez-Granda, 2014, Candès and Fernandez-Granda, 2013] among others), our bounds rely on the existence of interpolating functions called "certificates" (see Assumptions 2.6.1 and 2.6.2) instead of relying on compatibility conditions or Restricted Eigenvalue conditions. We give an explicit way to construct such functions in the spirit of [Poon et al., 2021]. We show in this chapter that such functions can be constructed provided the non-linear parameters belonging to Θ are well separated with respect to a Riemannian metric $\mathfrak{d}_T$ (defined in Section 2.4.1) associated to the kernel $\mathcal{K}_T(\theta, \theta') = \langle \phi_T(\theta), \phi_T(\theta') \rangle_T$. See Remark 2.8.2 for comments on the separation distance in the particular case of sparse spike deconvolution. The Riemannian metric appears naturally when it comes to tackle a wide variety of dictionaries. In addition, it leads to a lot of invariances in many quantities useful in the proofs. Typically, the Riemannian metrics $\mathfrak{d}_T$ and $\mathfrak{d}_T^{\mathrm{h}}$ associated respectively to the kernel $\mathcal{K}_T(\cdot, \cdot)$ and the warped kernel $\mathcal{K}_T^{\mathrm{h}} = \mathcal{K}_T(\mathrm{h}(\cdot), \mathrm{h}(\cdot))$ for some smooth enough diffeomorphism h are equal and we have $\mathfrak{d}_T(\theta, \theta') = \mathfrak{d}_T^{\mathrm{h}}(\mathrm{h}^{-1}(\theta), \mathrm{h}^{-1}(\theta'))$.

Our statistical results rely on tail bounds for suprema of Gaussian processes: following [Boyer et al., 2017], instead of using controls on $\|w_T\|_T$ as in the seminal works [Duval and Peyré, 2015, Poon et al., 2021], we used bounds, based on the noise structure from Assumption 2.1.1, on quantities of the form $\sup_{\Theta_T} \langle f(\theta), w_T \rangle_T$ for some H_T-valued functions f built from the dictionary $(\varphi_T(\theta), \theta \in \Theta)$ and its derivative. This approach is relevant as for some models the quantity $\|w_T\|_T$ may be very large, see for example the truncated white noise model from Section 2.1.2.3.

2.2 Main Results

Recall that we consider the model (2.2) that we can write in an equivalent way as, with $S^\star$ the support of the vector $\beta^\star$:

$$y = \sum_{j \in S^\star} \beta_j^\star \frac{\varphi_T(\theta_j^\star)}{\|\varphi_T(\theta_j^\star)\|_T} + w_T \quad \text{in } H_T.$$

The main theorem of this chapter gives the behavior of the prediction error with respect to: the decay rate of the noise variance Δ_T, the parameter $T \in \mathbb{N}$, the sparsity $s \in \mathbb{N}^*$, the upper bound on the number of components in the mixed signal K and the intrinsic noise level σ. We shall consider assumptions on the regularity of the dictionary φ_T, on the parameter space Θ_T on which the optimization is performed and on the noise w_T. Using the features φ_T we build a kernel $\mathcal{K}_T$ on the space of parameters Θ and an associated Riemannian metric $\mathfrak{d}_T$, see Section 2.4, which is the intrinsic metric, rather than the usual Euclidean metric. More assumptions are necessary on the closeness of the kernel $\mathcal{K}_T$ and its derivatives defined in (2.26) to a limit kernel $\mathcal{K}_\infty$ and its derivatives.

The theorem is stated assuming the existence of certificate functions, see Assumptions 2.6.1 and 2.6.2. Sufficient conditions for their existence are given later in Section 2.7, in which Propositions 2.7.4 and 2.7.5 show that the limit kernel $\mathcal{K}_\infty$ must be uniformly bounded and have concavity properties. In this case, the existence of certificates stands provided the underlying non-linear parameters to be estimated are sufficiently separated according to the Riemannian metric $\mathfrak{d}_T$, see Condition *(iii)* in Propositions 2.7.4 and 2.7.5.

In the following result the parameter set Θ_T is a one dimensional compact interval. We note $|\Theta_T|_{\mathfrak{d}_T}$ its length with respect to the Riemannian metric $\mathfrak{d}_T$ on Θ^2 associated to the kernel $\mathcal{K}_T$.

Theorem 2.2.1. *Assume we observe the random element y of H_T under the regression model (2.2) with unknown parameters $\beta^\star$ and $\vartheta^\star = (\theta_1^\star, \cdots, \theta_K^\star)$ a vector with entries in Θ_T, a compact interval of $\mathbb{R}$, such that:*

(i) ***Admissible noise:*** *The noise process w_T satisfies Assumption 2.1.1 for a noise level $\sigma > 0$ and a decay rate for the noise variance $\Delta_T > 0$.*

(ii) ***Regularity of the dictionary φ_T:*** *The dictionary function φ_T satisfies the smoothness conditions of Assumption 2.3.1. The function g_T defined in (2.12), satisfies the positivity condition of Assumption 2.3.2.*

(iii) ***Regularity of the limit kernel:*** *The kernel $\mathcal{K}_\infty$ and the functions g_∞ and h_∞, defined on an interval $\Theta_\infty \subset \Theta$, see (2.14) and (2.30), satisfy the smoothness conditions of Assumption 2.5.1.*

(iv) ***Proximity to the limit kernel:*** *The kernel $\mathcal{K}_T$ defined from the dictionary, see (2.26), is sufficiently close to the limit kernel $\mathcal{K}_\infty$ in the sense that Assumption 2.5.2 holds.*

(v) ***Existence of certificates:*** *The set of unknown parameters $\mathcal{Q}^\star = \{\theta_k^\star,\, k \in S^\star\}$, with $S^\star = \mathrm{Supp}(\beta^\star)$, satisfies Assumptions 2.6.1 and 2.6.2 with the same $r > 0$.*

Then, there exist finite positive constants $\mathcal{C}_0$, $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$ depending on the kernel $\mathcal{K}_\infty$ defined on Θ_∞ and on r such that for any $\tau > 1$ and a tuning parameter:

$$\kappa \geq \mathcal{C}_1 \sigma \sqrt{\Delta_T \log \tau},$$

we have the prediction error bound of the estimators $\hat{\beta}$ and $\hat{\vartheta}$ defined in (2.3) given by:

$$\left\| \hat{\beta} \Phi_T(\hat{\vartheta}) - \beta^\star \Phi_T(\vartheta^\star) \right\|_T \leq \mathcal{C}_0 \sqrt{s}\, \kappa, \tag{2.4}$$

with probability larger than $1 - \mathcal{C}_2 \left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau\sqrt{\log \tau}} \vee \frac{1}{\tau} \right)$. *Moreover, with the same probability, the difference of the* ℓ_1*-norms of* $\hat{\beta}$ *and* $\beta^\star$ *is bounded by:*

$$\left| \| \hat{\beta} \|_{\ell_1} - \| \beta^\star \|_{\ell_1} \right| \leq \mathcal{C}_3 \, \kappa \, s. \tag{2.5}$$

This result holds for both the continuous and discrete settings described in Section 2.1.2, covers a wide range of smooth dictionaries, and is proven under mild assumptions on the noise. We discuss in the next remark that the prediction error is, up to a logarithmic factor, almost optimal.

Remark 2.2.2 (Comparison with the Lasso estimator). Let us consider the model of Section 2.1.2.1 where the observation space is the Hilbert space $H_T = \mathbb{R}^T$ endowed with the Euclidean norm $\|\cdot\|_{\ell_2}$. The observation $y \in \mathbb{R}^T$ comes from the model (2.2) where the noise is a Gaussian vector with independent entries of variance σ^2. In this setting, the decay rate of the noise variance is fixed with $\Delta_T = 1$.

We first consider that the parameters $\vartheta^\star$ are known. In this case, the model becomes the classical high-dimensional regression model and the Lasso estimator $\hat{\beta}_L$ can be used to estimate $\beta^\star$ under coherence assumptions on the finite dictionary made of the rows of the matrix $\Phi^\star = \Phi_T(\vartheta^\star)$ (see [Bickel et al., 2009]). The behavior of the Lasso estimator has been studied in the literature and its prediction risk tends to zero at the rate:

$$\frac{1}{T} \| (\hat{\beta}_L - \beta^\star) \Phi^\star \|_{\ell_2}^2 = \mathcal{O}\left(\frac{\sigma^2 \, s \, \log(K)}{T} \right)$$

with high probability, larger than $1 - 1/K^\gamma$ for some positive constant $\gamma > 0$. Furthermore, in the case where $\beta^\star$ is an unknown s-sparse vector, $\vartheta^\star$ is known and $\Phi^\star$ verifies a coherence property, then the lower bounds of order $\sigma^2 \, s \, \log(K/s)/T$ in expected value can be deduced from the more general bounds for group sparsity in [Lounici et al., 2011] (see also [Raskutti et al., 2011]). The non-asymptotic prediction lower bounds for the prediction error given in [Raskutti et al., 2011] are:

$$\inf_{\hat{\beta}} \sup_{\beta^\star \, s-\text{sparse}} \mathbb{E}\left[\frac{1}{T} \| (\hat{\beta} - \beta^\star) \Phi^\star \|_{\ell_2}^2 \right] \geq C \cdot \frac{\sigma^2 \, s \, \log(K/s)}{T},$$

where the infinimum is taken over all the estimators $\hat{\beta}$ (square integrable measurable functions of the obervation y) and for some constant $C > 0$ free of s and T. When the parameters $\vartheta^\star$ are unknown, Theorem 2.2.1 gives an upper bound for the prediction risk which is, up to a logarithmic factor, almost the best rate we could achieve even knowing the non-linear parameters $\vartheta^\star$. Consider the estimators in (2.3) where the Riemannian diameter of the set Θ_T is bounded by a constant free of T (this is the case of Example 2.5.1 below). By squaring (2.4) and then dividing it by T, we obtain from Theorem 2.2.1 with $\kappa = \mathcal{C}_1 \sigma \sqrt{\Delta_T \log \tau}$ and $\tau = T^\gamma$ for some given $\gamma > 0$, that with high probability, larger than $1 - C/T^\gamma$:

$$\frac{1}{T} \left\| \hat{\beta} \Phi_T(\hat{\vartheta}) - \beta^\star \Phi_T(\vartheta^\star) \right\|_{\ell_2}^2 = \mathcal{O}\left(\frac{\sigma^2 \, s \, \log(T)}{T} \right). \tag{2.6}$$

Let us mention that [Tang et al., 2015] also obtained a similar prediction error (2.6) for the specific dictionary given by the complex exponential functions $(\varphi(\theta) : t \mapsto \mathrm{e}^{i2\pi t\theta}, \theta \in \Theta = [0, 2\pi])$; notice that the proof therein use the Parseval's identity for Fourier series as well as Markov-Bernstein type inequalities for trigonometric polynomials. Even if the structure of our proof is in the spirit of [Tang et al., 2015], our result is more general and does not rely on the convex setting of the BLasso approach.

Remark 2.2.3 (Proximity to the limit kernel). We comment on Condition *(iv)* on the proximity of the kernels $\mathcal{K}_T$ and $\mathcal{K}_\infty$, which also appears as Conditions *(iv)*-*(v)* in Proposition 2.7.4 (and similarly as Condition *(iv)* in Proposition 2.7.5).

In the examples of Sections 2.3.3.2 and 2.3.3.4 on translation or scaling model with a continuum of observations, the parameter T does not play any role in the definition of $\mathcal{K}_T$, so that one can take $\mathcal{K}_\infty$ equal to $\mathcal{K}_T$. In this case, the proximity conditions on the kernels are trivially satisfied.

In the continuation of Example 2.1.1, the example from Section 2.8 is devoted to the sparse spike deconvolution, that is, to a mixture of Gaussian translation invariant features observed in a discrete regression model on a regular grid of size T. In this case, we built a family of models $(H_T, \varphi_T, w_T, \Theta_T)$ with a dictionary φ_T which does not depend on T and such that the kernel $\mathcal{K}_T$ and its derivatives converge to $\mathcal{K}_\infty$ (and also ρ_T from (2.32) converges to 1). In this setting, the proximity condition of Theorem 2.2.1 holds for T large enough, say T larger than some T_0 which depends on $\mathcal{K}_\infty$, see Assumption 2.5.2. The existence of the certificates, see Propositions 2.7.4 and 2.7.5, also requires a proximity criterion which is achieved for T large enough, say T larger than some T_1 which depends on $\mathcal{K}_\infty$ and is increasing with the sparsity parameter s (see for example Condition (v) in Proposition 2.7.4).

Remark 2.2.4 (On the dimension K, the upper bound of the sparsity). We remark that neither the bound on the prediction error nor the probability on which the bound holds, depends on the upper bound K on the sparsity s. Therefore, the value of K can be taken arbitrarily large. It is not surprising that K does not have any impact on the bound since the optimisation problem (2.3) could be formulated without any bound on the sparsity. Indeed, the problem (2.3) can be embedded in an optimization problem over a space of measures following the literature on the BLasso introduced in [de Castro and Gamboa, 2012]. See also Remark 2.2.7.

The next theorem gives bounds on the differences between the parameters $\hat{\beta}$ given by the optimization problem (2.3) and the "true" parameters $\beta^\star$ for active features having their parameter $\hat{\theta}_\ell$ close, with respect to the Riemannian metric $\mathfrak{d}_T$, to a parameter $\theta^\star_k$, with k in $S^\star$. For $r > 0$ given by Assumptions 2.6.1 and 2.6.2, we define:

- The support of $\hat{\beta}$ given by the optimization problem (2.3): $\hat{S} = \mathrm{Supp}(\hat{\beta}) = \left\{\ell : \ \hat{\beta}_\ell \neq 0\right\}$.
- The near region $\tilde{S}(r)$ given by:

$$\tilde{S}(r) = \bigcup_{k \ \in S^\star} \tilde{S}_k(r) \quad \text{where} \quad \tilde{S}_k(r) = \left\{\ell \in \hat{S} : \mathfrak{d}_T(\hat{\theta}_\ell, \theta^\star_k) \leq r\right\},$$

  which corresponds to the set of indices ℓ in the support of $\hat{\beta}$ such that the corresponding parameter $\hat{\theta}_\ell$ is close to one of the true parameter $\theta^\star_k$, for some $k \in S^\star$.

The set $\hat{S} \backslash \tilde{S}(r)$ is also called the far region. Notice that the sets $\tilde{S}_k(r)$ with $k \in S^\star$ are pairwise disjoint under Assumption 2.6.1, and that they can be empty. In what follows, we use the convention $\sum_\emptyset = 0$.

Theorem 2.2.5. *We consider the model in Theorem 2.2.1 and suppose that Assumptions (i)-(v) therein hold. Then, there exist finite positive constants $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$, $\mathcal{C}_4$, $\mathcal{C}_5$ depending on $\mathcal{K}_\infty$ defined on Θ_∞ and on r such that for any $\tau > 1$ and a tuning parameter:*

$$\kappa \geq \mathcal{C}_1 \sigma \sqrt{\Delta_T \log \tau}$$

the estimator $\hat{\beta}$ defined in (2.3) satisfies the following bounds with probability larger than $1 - \mathcal{C}_2 \left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau\sqrt{\log \tau}} \vee \frac{1}{\tau}\right)$:

$$\sum_{k \in S^\star} \left| |\beta^\star_k| - \sum_{\ell \in \tilde{S}_k(r)} |\hat{\beta}_\ell| \right| \leq \mathcal{C}_3 \, \kappa \, s, \quad \sum_{k \in S^\star} \left| \beta^\star_k - \sum_{\ell \in \tilde{S}_k(r)} \hat{\beta}_\ell \right| \leq \mathcal{C}_4 \, \kappa \, s \quad \text{and} \quad \left\| \hat{\beta}_{\tilde{S}(r)^c} \right\|_{\ell_1} \leq \mathcal{C}_5 \, \kappa \, s, \tag{2.7}$$

where for a subset S of $\mathcal{I} = \{1, \cdots, K\}$, the set S^c denotes the complementary set of S in $\mathcal{I}$, that is $\mathcal{I} \setminus S$.

Notice that each linear parameter $\beta_k^\star$ can be estimated by the sum of several linear coefficients $\hat{\beta}_\ell$ with $\ell \in \{1, \cdots, K\}$. The first two inequalities in (2.7) show that there can be no compensation between the estimators $\hat{\beta}_\ell$ that approximate the same $\beta_k^\star$ with $k \in S^\star$, meaning that there can be no large values of $\hat{\beta}_\ell$ having different signs that sum up to a possibly small (in absolute value) true $\beta_k^\star$. The second inequality in (2.7) gives the estimation rate of the linear parameters $\beta_k^\star$ with $k \in S^\star$. The last bound in (2.7) basically means that when an estimation $\hat{\theta}_\ell$ with $\ell \in \{1, \cdots, K\}$ is far from any parameter $\theta_k^\star$ with $k \in S^\star$, that is at a distance greater than r, the associated parameters $\hat{\beta}_\ell$ drop to zero if the tuning parameter κ is taken equal to its lower bound and the decay rate of the noise variance Δ_T drops to zero. Therefore, the contribution of the parameters $\hat{\theta}_\ell$ in the far region, that are not in $\tilde{S}(r)$, will drop to zero as well.

Remark 2.2.6 (Estimation rate for $\theta_k^\star$ with $k \in S^\star$). Under the assumptions of Theorem 2.2.5, we also have, with probability larger than $1 - \mathcal{C}_2 \left(\frac{|\Theta_T| \mathfrak{d}_T}{\tau \sqrt{\log \tau}} \vee \frac{1}{\tau} \right)$, the bound:

$$\sum_{k \in S^\star} \sum_{\ell \in \tilde{S}_k(r)} \left| \hat{\beta}_\ell \right| \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)^2 \leq \mathcal{C}_6 \, \kappa \, s. \tag{2.8}$$

This gives an estimation rate for the parameters $\theta_k^\star$ with $k \in S^\star$ when at least one estimator $\hat{\theta}_\ell$ given by the optimization problem (2.3) belongs to the near region $\tilde{S}_k(r)$, which is the Riemannian ball centered at $\theta_k^\star$ with radius r.

Remark 2.2.7 (Again on the dimension K). As in Theorem 2.2.1, we remark that neither the bounds nor the probability of the event on which the bounds hold depend on the upper bound K on the sparsity s.

If the optimization on ϑ in (2.3) is performed over a subset of Θ_T in which the coordinates of the considered vectors are at a distance greater than $2r$ pairwise with respect to the Riemannian metric $\mathfrak{d}_T$, then the sets $\tilde{S}_k(r)$ contain at most one element. However, by doing so, we introduce an upper bound on the dimension K whereas in Theorem 2.2.1 the dimension K can be arbitrarily large. Indeed, Θ_T is a compact set and therefore contains a finite number of balls of size $2r$.

Outline of the chapter. In Section 2.3, we give the definition of the kernel $\mathcal{K}_T$ measuring the correlation between two elements in the continuous dictionary and we present the regularity assumptions on the function φ_T. Section 2.4 introduces the Riemannian geometry framework useful in our context. Section 2.5 defines the convergence (or closeness condition) of kernels $\mathcal{K}_T$ towards a limit kernel $\mathcal{K}_\infty$. Then, we require properties on the limit kernel $\mathcal{K}_\infty$ and propagate them to the kernels $\mathcal{K}_T$ thanks to this convergence. In Section 2.6, we present the assumptions on the existence of the so-called certificate functions used to state Theorems 2.2.1 and 2.2.5. We give sufficient conditions for the existence of certificate functions in Section 2.7. The example of sparse deconvolution in our regression model is fully detailed in Section 2.8. Then, Sections 2.9.1 and 2.9.2 are dedicated to the proofs of Theorems 2.2.1 and 2.2.5. The proofs of existence and explicit constructions of the certificates are detailed in Section 2.10.

2.3 Dictionary of features

We recall in the next section some basic results on the Fréchet derivative and the Bochner integral. Then, we present the regularity assumptions on the features $(\varphi_T(\theta), \theta \in \Theta)$ we shall consider.

2.3.1 The Fréchet derivative and the Bochner integral

The Fréchet derivative and Bochner integrals are defined for Banach space valued functions, but we shall only consider the case of Hilbert space valued functions.

Let $(H, \langle \cdot, \cdot \rangle)$ be an Hilbert space and let Θ be an interval of $\mathbb{R}$. We note $\|\cdot\|$ the norm associated to the scalar product. A function f from Θ to H is Fréchet differentiable at $\theta \in \Theta$ if it is continuous at θ and there exists an element $\partial_\theta f \in H$ such that:

$$\lim_{h \to 0;\, \theta + h \in \Theta} \left\| \frac{f(\theta + h) - f(\theta)}{h} - \partial_\theta f(\theta) \right\| = 0.$$

The derivative of f is the function $\partial_\theta f : \theta \mapsto \partial_\theta f(\theta)$ defined on Θ when it exists. We also define by recurrence the derivative $\partial_\theta^i f$ of order $i \in \mathbb{N}^*$ of f as the derivative of $\partial_\theta^{i-1} f$, with the convention that $\partial_\theta^0 f = f$, and say that f is of class $\mathcal{C}^i$ if the derivatives $\partial_\theta^j f$ exist and are continuous on Θ for $j \in \{0, \dots, i\}$. The standard differentiating rules for composition, addition and multiplication apply to the Fréchet derivative. We refer to [Lang, 1993] for a complete presentation of the subject. By definition, if f is differentiable at $\theta \in \Theta$, then we have for all $g \in H$ that:

$$\partial_\theta \langle f(\theta), g \rangle = \langle \partial_\theta f(\theta), g \rangle . \tag{2.9}$$

The Bochner integral extends the Lebesgue integral. We refer to [Arendt et al., 2011, Chapter 1] and [Aliprantis and Border, 2006, Section 11.8] for further details on the Bochner integral. We endow the interval $\Theta \subset \mathbb{R}$ with its usual Borel sigma field inherited from the Borel sigma field on $\mathbb{R}$ and a measure μ. A function f from Θ to H is strongly measurable if it is the limit of simple functions or equivalently, see [Aliprantis and Border, 2006, Lemma 11.37], if the map $\theta \mapsto \langle f(\theta), g \rangle$ is measurable for all $g \in H$ and $f(\theta)$ lies for μ-almost every $\theta \in \Theta$ in a closed separable subspace of H. In particular if the function f is continuous, then it is strongly measurable, see [Arendt et al., 2011, Corollary 1.1.2]. If f is strongly measurable, then the norm $\|f\|$ is a measurable function from Θ to $\mathbb{R}$, see [Aliprantis and Border, 2006, Lemma 11.39]. Then a function f defined on Θ (endowed with the Lebesgue measure) is Bochner integrable if and only if it is strongly measurable and if $\|f\|$ is integrable; in which case, we have $\|\int f(\theta) \, \mathrm{d}\theta\| \leq \int \|f(\theta)\| \, \mathrm{d}\theta$, see [Aliprantis and Border, 2006, Theorem 11.44] (which is easily extended from finite measure to σ-finite measure, see also [Arendt et al., 2011, Theorem 1.1.4] in this direction). We remark that the fundamental theorem of calculus is still valid in this framework, see [Arendt et al., 2011, Proposition 1.2.2]. In particular, if f is continuous and Bochner integrable on Θ and $\theta_0 \in \Theta$, then, we have:

$$F'(\theta) = f(\theta) \quad \text{where} \quad F(\theta) = \int_{\theta_0}^{\theta} f(q) \, \mathrm{d}q. \tag{2.10}$$

As a particular case of [Aliprantis and Border, 2006, Lemma 11.45], if f is Bochner integrable on Θ, then for all $g \in H$, we have that:

$$\int_\Theta \langle f(\theta), g \rangle \, \mathrm{d}\theta = \langle \int_\Theta f(\theta) \, \mathrm{d}\theta, g \rangle.$$

2.3.2 Assumptions on the regularity of the features

Let $T \in \mathbb{N}$ be fixed. We consider the Hilbert space $(H_T, \langle \cdot, \cdot \rangle_T)$ and the features $(\varphi_T(\theta), \theta \in \Theta)$ which are elements of H_T. We shall consider the following regularity assumptions on the features.

Assumption 2.3.1 (Smoothness of φ_T)**.** *We assume that the function $\varphi_T : \Theta \to H_T$ is of class $\mathcal{C}^3$ and $\|\varphi_T(\theta)\|_T > 0$ on Θ.*

Recall $\phi_T = \varphi_T / \|\varphi_T\|_T$ from (2.1) and notice that ϕ_T, and thus Φ_T, are continuous functions. Under Assumption 2.3.1, elementary calculations using (2.9) give:

$$\partial_\theta \phi_T(\theta) = \frac{\partial_\theta \varphi_T(\theta)}{\|\varphi_T(\theta)\|_T} - \frac{\varphi_T(\theta) \, \langle \varphi_T(\theta), \partial_\theta \varphi_T(\theta) \rangle_T}{\|\varphi_T(\theta)\|_T^3}, \tag{2.11}$$

and thus, we deduce that the function $g_T : \Theta \mapsto \mathbb{R}_+$ defined by:

$$g_T(\theta) = \|\partial_\theta \phi_T(\theta)\|_T^2 \tag{2.12}$$

is well defined and continuous.

We shall consider the following non-degeneracy assumption on the features.

Assumption 2.3.2 (Positivity of g_T)**.** *Assumption 2.3.1 holds and we have $g_T > 0$ on Θ.*

Even if Assumption 2.3.2 requires Assumption 2.3.1, in the following we shall stress when Assumption 2.3.1 is in force.

The next lemma gives a sufficient condition on φ_T for Assumption 2.3.2 to hold.

Lemma 2.3.1 (On the positivity of g_T)**.** *Suppose Assumption 2.3.1 holds. If the elements $\varphi_T(\theta)$ and $\partial_\theta \varphi_T(\theta)$ of H_T are linearly independent for all $\theta \in \Theta$ and $\|\partial_\theta \varphi_T(\theta)\|_T > 0$ for all $\theta \in \Theta$, then g_T is positive on Θ.*

Proof. For simplicity, we remove the subscript T, and for example write simply $\phi = \varphi/\|\varphi\|$. Recall that by Assumption 2.3.1 we have $\|\varphi(\theta)\| > 0$. Assume there exists $\theta \in \Theta$ such that $g(\theta) = 0$, that is $\partial_\theta \phi(\theta) = 0$. Since $\|\varphi(\theta)\| > 0$, we deduce from (2.11) that $\partial_\theta \varphi(\theta)\|\varphi(\theta)\|^2 - \varphi(\theta)\langle \varphi(\theta), \partial_\theta \varphi(\theta)\rangle = 0$. Then use that by assumption $\partial_\theta \varphi(\theta) \neq 0$ and $\|\varphi(\theta)\| > 0$, to get that $\varphi(\theta)$ and $\partial_\theta \varphi(\theta)$ are linearly dependent. In conclusion, we get that if $\varphi(\theta)$ and $\partial_\theta \varphi(\theta)$ are linearly independent, then $g(\theta) > 0$. □

2.3.3 Examples of regular features

The aim of this section of examples is to stress that a large variety of dictionaries of features and type of parameters verify Assumptions 2.3.1 and 2.3.2.

2.3.3.1 Translation model with observations on a finite grid

This model is an extension of Example 2.1.1 in the spirit of Section 2.1.2.2. Let $t_1 < \cdots < t_T$ be a grid on $\mathbb{R}$ of size $T \in \mathbb{N}$, λ_T an atomic measure whose support is the grid, and $H_T = L^2(\lambda_T)$. Consider the translation invariant dictionary:

$$(\varphi_T(\theta) = k(\cdot - \theta),\ \theta \in \Theta), \tag{2.13}$$

with $\Theta = \mathbb{R}$ and k is a real-valued $\mathcal{C}^3$ function defined on $\mathbb{R}$. Notice the dictionary does not depend on T. We now consider usual choices for the function k.

For the Gaussian function $k(t) = \mathrm{e}^{-t^2/2}$ and the Cauchy function $k(t) = 1/(1+t^2)$, we get that Assumption 2.3.1 holds and, using Lemma 2.3.1 that Assumption 2.3.2 is also satisfied provided respectively $T \geq 2$ and $T \geq 3$.

For the Shannon scaling function $k(t) = \mathrm{sinc}(t) = \sin(\pi t)/(\pi t)$, Assumption 2.3.1 holds provided that $\lambda_T((a+\mathbb{Z})^c) > 0$ for all $a \in \mathbb{R}$, that is the grid is not a subset of $a+\mathbb{Z}^*$ for some $a \in \mathbb{R}$. There is no easy way to write conditions on the grid, based on the use of Lemma 2.3.1, for Assumption 2.3.2 to hold (let us mention that $T \geq 2$ and $\min_{1 \leq i \leq T-1}(t_{i+1} - t_i) < 1/2$ is a sufficient condition for Assumption 2.3.2 to hold).

Eventually notice that the Laplace function $k(t) = \mathrm{e}^{-|t|}$ is not smooth enough for Assumption 2.3.1 to hold.

2.3.3.2 Translation model with a continuum of observations

Let $T \in \mathbb{N}$ (which does not play a role here) and $H_T = L^2(\mathrm{Leb})$, where Leb is the Lebesgue measure on $\mathbb{R}$. In this framework, the observation y defined in (2.2) is a continuum of observations. Consider the translation invariant dictionaries from Section 2.3.3.1, where k is either the Gaussian, the Cauchy or the Shannon scaling function. Notice that the Hilbert

space and the dictionary do not depend on T. Then, it is easy to check that Assumptions 2.3.1 and 2.3.2 hold.

We see that this model, which can be seen as a continuous approximation (or limit) of the discrete models from Section 2.3.3.1 when T therein is large, is easier to handle than the corresponding discrete models.

2.3.3.3 Translation model with a varying scaling parameter

Let $T \in \mathbb{N}$, $H_T = L^2(\mathrm{Leb})$, where Leb is the Lebesgue measure on $\mathbb{R}$, and consider the translation invariant dictionary scaled by $\overline{\sigma}_T > 0$ given by:

$$(\varphi_T(\theta) = k(\overline{\sigma}_T^{-1}(\cdot - \theta)),\ \theta \in \Theta),$$

with $\Theta = \mathbb{R}$ and k is a real-valued $\mathcal{C}^3$ function defined on $\mathbb{R}$. Contrary to Section 2.3.3.2, the features depend on T. Suppose that k is the Shannon scaling function (see Section 2.3.3.1) and consider the vector sub-space V_T given by the closure in H_T of the vector space spanned by the dictionary. According to [Mallat, 2009, Theorem 3.5], the set V_T is the subset of H_T of all functions whose Fourier transform support is a subset of $[-\pi/\overline{\sigma}_T, \pi/\overline{\sigma}_T]$. Suppose that the sequence $(\overline{\sigma}_T, T \in \mathbb{N})$ is decreasing to 0. Then the sequence $(V_T, T \in \mathbb{N})$ is increasing and $\overline{\bigcup_{T\in\mathbb{N}} V_T} = H_T$. This model provides an example of translation models with possibly varying, but known, scaling parameter $\overline{\sigma}_T$.

2.3.3.4 Scaling exponential model

Let $T \in \mathbb{N}$ (which does not play a role here), $H_T = L^2(\mathrm{Leb})$, where Leb is the Lebesgue measure on $\mathbb{R}_+$, and consider the scale invariant dictionary given by:

$$(\varphi_T(\theta) = k(\theta\cdot),\ \theta \in \Theta),$$

with $\Theta = \mathbb{R}_+^*$ and the exponential function $k : t \mapsto \mathrm{e}^{-t}$. This dictionary is used for example in fluorescence microscopy (see [Denoyelle et al., 2020]). Clearly Assumption 2.3.1 holds as well as Assumption 2.3.2 as $g_T(\theta) = 1/(4\theta^2)$.

2.4 A Riemannian metric on the set of parameters

2.4.1 On the Riemannian metric in dimension one

Recall Θ is an interval of $\mathbb{R}$. We call kernel a real-valued function defined on Θ^2. Let $\mathcal{K}$ be a symmetric kernel of class $\mathcal{C}^2$ such that the function $g_{\mathcal{K}}$ defined on Θ by:

$$g_{\mathcal{K}}(\theta) = \partial^2_{x,y}\mathcal{K}(\theta, \theta) \tag{2.14}$$

is positive and locally bounded, where ∂_x (resp. ∂_y) denotes the usual derivative with respect to the first (resp. second) variable. Following [Poon et al., 2021], we define an intrinsic Riemannian metric, denoted $\mathfrak{d}_{\mathcal{K}}$, on the parameter set Θ using the function $g_{\mathcal{K}}$. One of the motivation to use the Riemannian metric is to work with intrinsic quantities related to the parameters which are invariant by reparametrization, such as the diameter of (subsets of) Θ. Since Θ is one-dimensional and connected, the Riemannian metric $\mathfrak{d}_{\mathcal{K}}(\theta, \theta')$ between $\theta, \theta' \in \Theta$ reduces to:

$$\mathfrak{d}_{\mathcal{K}}(\theta, \theta') = |G_{\mathcal{K}}(\theta) - G_{\mathcal{K}}(\theta')|, \tag{2.15}$$

where $G_{\mathcal{K}}$ is a primitive of $\sqrt{g_{\mathcal{K}}}$.

Remark 2.4.1. We refer to [Lee, 2018] and [Sakai, 1996] for a general presentation on Riemannian manifolds, and we give an immediate application in dimension one which entails in particular (2.15). Let Θ be a manifold (of dimension one). A path $\gamma : [0, 1] \to \Theta$ is an

admissible path if it is continuous, piecewise continuously differentiable with non-vanishing derivative. Its length is given by $\mathcal{L}_{\mathcal{K}}(\gamma) = \int_0^1 |\dot{\gamma}_s| \sqrt{g_{\mathcal{K}}(\gamma_s)}\,\mathrm{d}s$, where $|\dot{\gamma}_s|$ is seen as the norm of the vector $\dot{\gamma}_s$ in the tangent space, and the scalar product on the tangent space at $\theta \in \Theta$ is given by $(u, v) \mapsto \langle u, g_{\mathcal{K}}(\theta)v\rangle$ with $\langle\cdot,\cdot\rangle$ the usual Euclidean scalar product. (In our case, the tangent vector space is $\mathbb{R}$ and the Euclidean scalar product reduces to the usual product). The Riemannian metric $\mathfrak{d}_{\mathcal{K}}$ between θ, θ' in Θ is then defined by:

$$\mathfrak{d}_{\mathcal{K}}(\theta, \theta') = \inf_{\gamma} \quad \mathcal{L}_{\mathcal{K}}(\gamma), \tag{2.16}$$

where the infinimum is taken over the admissible paths γ such that $\gamma_0 = \theta$ and $\gamma_1 = \theta'$. It is not hard to see that γ is a minimizing path, that is, $\mathfrak{d}_{\mathcal{K}}(\theta, \theta') = \mathcal{L}_{\mathcal{K}}(\gamma)$, if and only if γ is monotone (and thus $\gamma_s \in [\theta \wedge \theta', \theta \vee \theta']$ for all $s \in [0, 1]$). This is equivalent to say that the sign of $\dot{\gamma}_s$ is constant. Assume that $g_{\mathcal{K}}$ is of class $\mathcal{C}^1$. The path γ is a geodesic if it is smooth with zero acceleration, that is, in dimension one for all $s \in (0, 1)$:

$$\ddot{\gamma}_s + \frac{1}{2}\frac{g'_{\mathcal{K}}(\gamma_s)}{g_{\mathcal{K}}(\gamma_s)}\dot{\gamma}_s^2 = 0.$$

This is equivalent to $s \mapsto \dot{\gamma}_s \sqrt{g_{\mathcal{K}}(\gamma_s)}$ being constant, which implies that the geodesic is a minimizing path.

We now derive the equation of the geodesic path when $g_{\mathcal{K}}$ is of class $\mathcal{C}^1$. Recall $G_{\mathcal{K}}$ denotes the primitive of $\sqrt{g_{\mathcal{K}}}$. It is continuous increasing and thus induces a one-to-one map from Θ to its image. Set $a = G_{\mathcal{K}}(\theta)$ and $b = G_{\mathcal{K}}(\theta') - G_{\mathcal{K}}(\theta)$, so that the path $\gamma : [0, 1] \to \Theta$ defined by $\gamma_s = G_{\mathcal{K}}^{-1}(a + bs)$ is a geodesic and minimizing path from θ to θ' with $\mathcal{L}_{\mathcal{K}}(\gamma) = \mathfrak{d}_{\mathcal{K}}(\theta, \theta')$.

Following [Poon et al., 2021], we introduce the covariant derivatives, see [Absil et al., 2008, Sections 3.6 and 5.6], which have elementary expressions as the set of parameters Θ is one-dimensional. For a smooth function f defined on Θ and taking values in an Hilbert space, say H, the covariant derivative $D_{i;\mathcal{K}}[f]$ of order $i \in \mathbb{N}$ is defined recursively by $D_{0;\mathcal{K}}[f] = f$ and for $i \in \mathbb{N}$, assuming that $g_{\mathcal{K}}$ is of class $\mathcal{C}^i$, and $\theta \in \Theta$:

$$D_{i+1;\mathcal{K}}[f](\theta) = g_{\mathcal{K}}(\theta)^{\frac{i}{2}}\partial_\theta\left(\frac{D_{i;\mathcal{K}}[f](\theta)}{g_{\mathcal{K}}(\theta)^{\frac{i}{2}}}\right). \tag{2.17}$$

In particular, we have for $f \in \mathcal{C}^2(\Theta, H)$ (and assuming that $g_{\mathcal{K}}$ is of class $\mathcal{C}^1$ for the last equality) that:

$$D_{0;\mathcal{K}}[f] = f, \quad D_{1;\mathcal{K}}[f] = \partial_\theta f, \quad D_{2;\mathcal{K}}[f] = \partial_\theta^2 f - \frac{1}{2}\frac{g'_{\mathcal{K}}}{g_{\mathcal{K}}}\partial_\theta f. \tag{2.18}$$

We shall also consider the following modification of the covariant derivative, for $i \in \mathbb{N}$:

$$\tilde{D}_{i;\mathcal{K}}[f](\theta) = g_{\mathcal{K}}(\theta)^{-i/2}\, D_{i;\mathcal{K}}[f](\theta). \tag{2.19}$$

We have $\tilde{D}_{0;\mathcal{K}}[f] = f$, and we deduce from (2.17) that for $i \in \mathbb{N}^*$, assuming that $g_{\mathcal{K}}$ is of class $\mathcal{C}^i$:

$$\tilde{D}_{i;\mathcal{K}} = \tilde{D}_{1;\mathcal{K}} \circ \tilde{D}_{i-1;\mathcal{K}} = \left(\tilde{D}_{1;\mathcal{K}}\right)^i, \tag{2.20}$$

so that $\tilde{D}_{1;\mathcal{K}}$ can be seen as a derivative operator.

We now give an elementary variant of the Taylor-Lagrange expansion using the previously defined Riemannian metric and covariant derivatives. Its proof can be found in the Appendix, Section 2.11.2.1.

Lemma 2.4.2. *Assume $g_{\mathcal{K}}$ is positive and of class $\mathcal{C}^1$. Let f be a function defined on Θ taking values in an Hilbert space of class $\mathcal{C}^2$. Setting $f^{[i]} = \tilde{D}_{i;\mathcal{K}}[f]$ for $i \in \{1,2\}$, we have that for all $\theta, \theta_0 \in \Theta$:*

$$f(\theta) = f(\theta_0) + \operatorname{sign}(\theta - \theta_0)\, \mathfrak{d}_{\mathcal{K}}(\theta, \theta_0)\, f^{[1]}(\theta_0) + \mathfrak{d}_{\mathcal{K}}(\theta, \theta_0)^2 \int_0^1 (1-t) f^{[2]}(\gamma_t)\, \mathrm{d}t, \tag{2.21}$$

where γ is a geodesic path such that $\gamma_0 = \theta_0$, $\gamma_1 = \theta$ (and thus $\mathfrak{d}_{\mathcal{K}}(\theta, \theta_0) = \mathcal{L}_{\mathcal{K}}(\gamma)$).

For a real-valued function F defined on Θ^2, we say that F is of class $\mathcal{C}^{0,0}$ on Θ^2 if it is continuous on Θ^2, and of class $\mathcal{C}^{i,j}$ on Θ^2, with $i, j \in \mathbb{N}$, as soon as: F is of class $\mathcal{C}^{0,0}$, and if $i \geq 1$ then the function $\theta \mapsto F(\theta, \theta')$ is of class $\mathcal{C}^i$ on Θ and its derivative $\partial_x F$ is of class $\mathcal{C}^{i-1,j}$ on Θ^2, and if $j \geq 1$ the function $\theta' \mapsto F(\theta, \theta')$ is of class $\mathcal{C}^j$ on Θ and its derivative $\partial_y F$ is of class $\mathcal{C}^{i,j-1}$ on Θ^2. For a real-valued symmetric function F defined on Θ^2 of class $\mathcal{C}^{i,j}$, we define the covariant derivatives $D_{i,j;\mathcal{K}}[F]$ of order $(i,j) \in \mathbb{N}^2$ recursively by $D_{0,0;\mathcal{K}}[F] = F$ and for $i, j \in \mathbb{N}$, assuming that $g_{\mathcal{K}}$ is of class $\mathcal{C}^{\max(i,j)}$, and $\theta, \theta' \in \Theta$:

$$D_{i+1,j;\mathcal{K}}[F](\theta, \theta') = g_{\mathcal{K}}(\theta)^{\frac{i}{2}} \partial_\theta \left(\frac{D_{i,j;\mathcal{K}}[F](\theta, \theta')}{g_{\mathcal{K}}(\theta)^{\frac{i}{2}}} \right) \quad \text{and} \quad D_{i,j;\mathcal{K}}[F](\theta, \theta') = D_{j,i;\mathcal{K}}[F](\theta', \theta). \tag{2.22}$$

In particular, we have $D_{0,0;\mathcal{K}}[F] = F$, $D_{1,0;\mathcal{K}} = \partial_x F$, $D_{0,1;\mathcal{K}} = \partial_y F$ and $D_{1,1;\mathcal{K}} = \partial^2_{xy} F$. We shall also consider the following modification of the covariant derivative, for $i, j \in \mathbb{N}$:

$$\tilde{D}_{i,j;\mathcal{K}}[F](\theta, \theta') = \frac{D_{i,j;\mathcal{K}}[F](\theta, \theta')}{g_{\mathcal{K}}(\theta)^{i/2}\, g_{\mathcal{K}}(\theta')^{j/2}}. \tag{2.23}$$

We have $\tilde{D}_{1,0;\mathcal{K}} \circ \tilde{D}_{0,1;\mathcal{K}} = \tilde{D}_{0,1;\mathcal{K}} \circ \tilde{D}_{1,0;\mathcal{K}}$ and for $i, j \in \mathbb{N}$, assuming that $g_{\mathcal{K}}$ is of class $\mathcal{C}^{\max(i,j)}$:

$$\tilde{D}_{i,j;\mathcal{K}} = \left(\tilde{D}_{1,0;\mathcal{K}}\right)^i \circ \left(\tilde{D}_{0,1;\mathcal{K}}\right)^j.$$

For $i, j \in \mathbb{N}$, if $\mathcal{K}$ is of class $\mathcal{C}^{i \vee 1, j \vee 1}$, we consider the real-valued function defined on Θ^2 by:

$$\mathcal{K}^{[i,j]} = \tilde{D}_{i,j;\mathcal{K}}[\mathcal{K}]. \tag{2.24}$$

In particular, since $\mathcal{K}$ is of class $\mathcal{C}^2$, we have:

$$\mathcal{K}^{[0,0]} = \mathcal{K} \quad \text{and} \quad \mathcal{K}^{[1,1]}(\theta, \theta) = 1. \tag{2.25}$$

2.4.2 The kernel and the Riemannian metric associated to the dictionary of features

Let $T \in \mathbb{N}$ be fixed and assume that Assumption 2.3.2 holds. We define the kernel $\mathcal{K}_T$ on Θ^2 by:

$$\mathcal{K}_T(\theta, \theta') = \langle \phi_T(\theta), \phi_T(\theta') \rangle_T = \frac{\langle \varphi_T(\theta), \varphi_T(\theta') \rangle_T}{\|\varphi_T(\theta)\|_T \|\varphi_T(\theta')\|_T}, \tag{2.26}$$

where we recall that $\phi_T = \varphi_T / \|\varphi_T\|_T$. When considering the kernel $\mathcal{K}_T$, we shall write g_T for $g_{\mathcal{K}_T}$, and similarly we shall use the notations $\tilde{D}_{i;T}$ and $\tilde{D}_{i,j;T}$ instead of $\tilde{D}_{i;\mathcal{K}_T}$ and $\tilde{D}_{i,j;\mathcal{K}_T}$. Recall the derivatives of the kernel $\mathcal{K}_T$ defined by (2.24). The next lemma insures in particular that the two definitions of g_T given by (2.12) and (2.14) are consistent, that is:

$$g_T(\theta) = \partial^2_{xy} \mathcal{K}_T(\theta, \theta) = \|\partial_\theta \phi_T(\theta)\|^2_T. \tag{2.27}$$

The proof of the next lemma can be found in the Appendix, Section 2.11.2.1.

Lemma 2.4.3. *Let $T \in \mathbb{N}$ be fixed and assume that Assumptions 2.3.1 and 2.3.2 hold. Then, the symmetric kernel $\mathcal{K}_T$ is of class $\mathcal{C}^{3,3}$ on Θ^2 and for $i,j \in \{0,\dots,3\}$ and $\theta,\theta' \in \Theta$, we have:*

$$\mathcal{K}_T^{[i,j]}(\theta,\theta') = \langle \tilde{D}_{i;T}[\phi_T](\theta), \tilde{D}_{j;T}[\phi_T](\theta')\rangle_T. \tag{2.28}$$

We also have:

$$\sup_{\Theta^2} |\mathcal{K}_T^{[0,0]}| \le 1, \quad \mathcal{K}_T^{[0,0]}(\theta,\theta) = 1, \quad \mathcal{K}_T^{[1,0]}(\theta,\theta) = 0, \quad \mathcal{K}_T^{[2,0]}(\theta,\theta) = -1 \quad \text{and} \quad \mathcal{K}_T^{[2,1]}(\theta,\theta) = 0. \tag{2.29}$$

2.5 Approximating the kernel associated to the dictionary

In the section we detail the assumptions guaranteeing the approximation of the kernel $\mathcal{K}_T$ (which is usually difficult to compute) by a kernel $\mathcal{K}_\infty$ (which is easier to handle). Both kernels are defined on Θ^2, however, we shall qualify the approximation of $\mathcal{K}_T$ by $\mathcal{K}_\infty$ and properties of $\mathcal{K}_\infty$ on subsets of Θ, respectively Θ_T (which will be a compact interval) and Θ_∞ (which will be an interval possibly unbounded). We use notations from Section 2.4 and recall the definition of $g_\mathcal{K}$, resp. $\mathcal{K}^{[i,j]}$, given in (2.14), resp. in (2.24). Assuming the kernel $\mathcal{K}$ is of class $\mathcal{C}^{3,3}$ and using the notation (2.24), we also set for $\theta \in \Theta$:

$$h_\mathcal{K}(\theta) = \mathcal{K}^{[3,3]}(\theta,\theta). \tag{2.30}$$

For simplicity, for an expression A we write A_* for $A_{\mathcal{K}_*}$ where $*$ is equal to T or ∞. We first give a regularity assumption on the kernel $\mathcal{K}_\infty$.

Assumption 2.5.1 (Properties of the asymptotic kernel $\mathcal{K}_\infty$ and function h_∞)**.** *The symmetric kernel $\mathcal{K}_\infty$ defined on Θ^2 is of class $\mathcal{C}^{3,3}$, the function g_∞ defined by (2.14) on Θ is positive and locally bounded (as well as of class $\mathcal{C}^2$), and we have $\mathcal{K}_\infty(\theta,\theta) = -\mathcal{K}_\infty^{[2,0]}(\theta,\theta) = 1$ for $\theta \in \Theta$. The set $\Theta_\infty \subseteq \Theta$ is an interval and we have for all $i,j \in \{0,1,2\}$:*

$$m_g := \inf_{\Theta_\infty} g_\infty > 0, \quad L_3 := \sup_{\Theta_\infty} h_\infty < +\infty, \quad \text{and} \quad L_{i,j} := \sup_{\Theta_\infty^2} |\mathcal{K}_\infty^{[i,j]}| < +\infty. \tag{2.31}$$

Since Θ_T is compact, under Assumptions 2.3.2 and 2.5.1, we deduce that the constant ρ_T below is positive and finite, where:

$$\rho_T = \max\left(\sup_{\Theta_T} \sqrt{\frac{g_T}{g_\infty}}, \sup_{\Theta_T} \sqrt{\frac{g_\infty}{g_T}}\right). \tag{2.32}$$

From the definition of the Riemannian metric given in (2.15) (see also (2.16)), we readily deduce that the metrics $\mathfrak{d}_T$ and $\mathfrak{d}_\infty$ are then strongly equivalent on Θ_T; more precisely we have that on Θ_T^2:

$$\frac{1}{\rho_T}\,\mathfrak{d}_\infty \le \mathfrak{d}_T \le \rho_T\,\mathfrak{d}_\infty. \tag{2.33}$$

We then give an assumption on the quality of approximation of $\mathcal{K}_T$ by $\mathcal{K}_\infty$. We set:

$$\mathcal{V}_T = \max(\mathcal{V}_T^{(1)}, \mathcal{V}_T^{(2)}) \quad \text{with} \quad \mathcal{V}_T^{(1)} = \max_{i,j\in\{0,1,2\}} \sup_{\Theta_T^2} |\mathcal{K}_T^{[i,j]} - \mathcal{K}_\infty^{[i,j]}| \quad \text{and} \quad \mathcal{V}_T^{(2)} = \sup_{\Theta_T} |h_T - h_\infty|. \tag{2.34}$$

Let us recall that Assumption 2.3.2 implies regularity conditions on $\mathcal{K}_T$, see Lemma 2.4.3.

Assumption 2.5.2 (Quality of the approximation)**.** *Let $T \in \mathbb{N}$ be fixed. Assumptions 2.3.2 and 2.5.1 hold, the interval $\Theta_T \subset \Theta_\infty$ is a compact interval, and we have:*

$$\mathcal{V}_T \le L_{2,2} \wedge L_3.$$

Notice that if Assumption 2.3.2 holds, then Assumptions 2.5.1 and 2.5.2 hold trivially when one takes $\mathcal{K}_\infty = \mathcal{K}_T$ and $\Theta_\infty = \Theta_T$; notice also that $\rho_T = 1$ in this case. In the next example, the sequence of kernels $(\mathcal{K}_T, T \in \mathbb{N})$ converges to the kernel $\mathcal{K}_\infty$ as T goes to infinity, so that Assumption 2.5.2 holds for T large enough.

Example 2.5.1. We consider the example from Example 2.1.1 with the framework of Section 2.1.2.2. We assume that the process y is a function defined on $[0,1]$ which, for $T \in \mathbb{N}^*$ is observed through the regular grid $\{t_{j,T} = j/T : 1 \leq j \leq T\}$. The process y is seen as an element of the Hilbert space $H_T = L^2(\lambda_T)$, with the probability measure $\lambda_T = \Delta_T \sum_{j=1}^T \delta_{t_{j,T}}$ on $[0,1]$ with $\Delta_T = 1/T$. Let Θ be a compact interval of $\mathbb{R}$ and set $\Theta_T = \Theta_\infty = \Theta$. Consider a dictionary $(\varphi(\theta), \theta \in \Theta)$ independent of T, that is, $\varphi_T = \varphi$ for all $T \in \mathbb{N}^*$, and assume that the function $(\theta, t) \mapsto \varphi(\theta)(t)$ is defined on $\Theta \times [0,1]$ and of class $\mathcal{C}^{3,0}$. Assume that the dictionary satisfies the regularity assumptions of Assumption 2.3.2.

Let Leb be the Lebesgue measure on $[0,1]$, so that $(\lambda_T, T \in \mathbb{N}^*)$ converges weakly to Leb. Then, define the kernel $\mathcal{K}_\infty$ by (2.26) with φ_T replaced by φ (as the dictionary does not depend on T) and the scalar product $\langle \cdot, \cdot \rangle_T$ by the usual scalar product on $L^2(\text{Leb})$. Thanks to Lemma 2.4.3, we deduce that Assumption 2.5.1 on the properties of $\mathcal{K}_\infty$ is satisfied. Using the weak convergence of $(\lambda_T, T \in \mathbb{N}^*)$ to Leb, we deduce that $\lim_{T\to\infty} \partial_x^i \partial_y^j \mathcal{K}_T = \partial_x^i \partial_y^j \mathcal{K}_\infty$ uniformly on $[0,1]^2$ for all $i, j \in \{0, \ldots, 3\}$. This implies that:

$$\lim_{T\to\infty} \mathcal{V}_T = 0 \quad \text{and} \quad \lim_{T\to\infty} \rho_T = 1.$$

Thus Assumption 2.5.2 holds for T large enough.

2.6 Certificates

In this section, we make assumptions on the existence of functions from Θ to $\mathbb{R}$ called certificates. These functions have interpolation properties that are corner stones in the proof of Theorem 2.2.1. The term "certificate" is inherited from the compressed sensing field were such functions were used to get rid of the Restricted Isometry Property condition (RIP) for exact recontruction of signals (see [Candes and Tao, 2005] for details on the RIP condition). In [Candès and Plan, 2011], the authors showed that is possible to reconstruct exactly a sparse signal from the observations of a finite number of Fourier coefficients by exhibiting a dual certificate. Many papers have followed this line of research since then (see *e.g.* [Candès and Fernandez-Granda, 2013, Candès and Fernandez-Granda, 2014, Duval and Peyré, 2015]).

In sparse linear models the bounds for prediction error are proved using RIP, Restricted Eigenvalue or compatibility conditions (see [Bickel et al., 2009, van de Geer, 2016]). Among these assumptions, the compatibility conditions are the less restrictive. Indeed, the authors of [van de Geer and Bühlmann, 2009] have shown that it is implied by both the RIP and the Restricted Eigenvalue. However, in many contexts even the weaker condition fails to hold. Typically the compatibility condition fails to hold in the context of super-resolution which aims at extracting the frequencies and amplitudes of a linear combination of complex exponentials from a small number of noisy time samples (see [Boyer et al., 2017]).

In the papers [Boyer et al., 2017] and [Tang et al., 2015], the authors achieve nearly optimal rates for the prediction error in the super-resolution framework using certificate functions. Their method and proof are however quite specific to complex exponentials and their certificates are trigonometric polynomials. The insightful paper of [Duval and Peyré, 2015] builds certificates in a quite general setting for a one dimensional parameter set Θ. In [De Castro et al., 2021], the authors exhibit certificate functions to deal with more general probability density models where Θ is multidimensional. However they are restricted to translation invariant dictionaries (2.13). The most general framework has been introduced in [Poon et al., 2021] where the Riemannian geometry is key to build in a natural way the so-called certificate functions. In fact a separation distance between the parameters to estimate is needed to build

certificates and the Euclidean metric yields overly pessimistic minimum separation condition. In what follows we introduce new certificates, called derivative certificates, in order to control the prediction error.

We consider the following assumption in the spirit of [Poon et al., 2021]. We consider the setting where T may be finite. Let $T \in \mathbb{N}$, H_T be an Hilbert space and $(\varphi_T(\theta), \theta \in \Theta)$ a dictionary satisfying Assumptions 2.3.1 and 2.3.2, so that the kernel $\mathcal{K}_T$ is of class $\mathcal{C}^{3,3}$ on Θ^2. Recall the Riemannian metric $\mathfrak{d}_{\mathcal{K}_T}$ associated to $\mathcal{K}_T$, which we simply denote by $\mathfrak{d}_T$. We define the closed ball centered at $\theta \in \Theta_T$ with radius r by:

$$\mathcal{B}_T(\theta, r) = \{\theta' \in \Theta_T,\, \mathfrak{d}_T(\theta, \theta') \leq r\} \subseteq \Theta_T.$$

Let $\mathcal{Q}^\star$ be a subset of Θ_T of cardinal s. For $r > 0$, the near region of $\mathcal{Q}^\star$ is the union of balls $\bigcup_{\theta^\star \in \mathcal{Q}^\star} \mathcal{B}_T(\theta^\star, r)$ and its far region is the complementary of the near region in Θ_T: $\Theta_T \setminus \bigcup_{\theta^\star \in \mathcal{Q}^\star} \mathcal{B}_T(\theta^\star, r)$. Sufficient conditions for the next assumption to hold are given in Section 2.7.

Assumption 2.6.1 (Interpolating certificate)**.** *Let $T \in \mathbb{N}$, $s \in \mathbb{N}^*$, $r > 0$ and $\mathcal{Q}^\star$ be a subset of Θ_T of cardinal s. Suppose Assumptions 2.3.1 and 2.3.2 on the dictionary $(\varphi_T(\theta), \theta \in \Theta)$, and Assumption 2.5.1 on the kernel $\mathcal{K}_\infty$, defined on Θ^2, hold. Suppose that $\mathfrak{d}_T(\theta, \theta') > 2r$ for all $\theta, \theta' \in \mathcal{Q}^\star \subset \Theta_T$, and that there exist finite positive constants C_N, C'_N, C_F, C_B, with $C_F < 1$, depending on r and $\mathcal{K}_\infty$ such that for any application $v : \mathcal{Q}^\star \to \{-1, 1\}$ there exists an element $p \in H_T$ satisfying:*

(i) For all $\theta^\star \in \mathcal{Q}^\star$ and $\theta \in \mathcal{B}_T(\theta^\star, r)$, we have $|\langle \phi_T(\theta), p\rangle_T| \leq 1 - C_N\, \mathfrak{d}_T(\theta^\star, \theta)^2$.

(ii) For all $\theta^\star \in \mathcal{Q}^\star$ and $\theta \in \mathcal{B}_T(\theta^\star, r)$, we have $|\langle \phi_T(\theta), p\rangle_T - v(\theta^\star)| \leq C'_N\, \mathfrak{d}_T(\theta^\star, \theta)^2$.

(iii) For all θ in Θ_T and $\theta \notin \bigcup_{\theta^\star \in \mathcal{Q}^\star} \mathcal{B}_T(\theta^\star, r)$ (far region), we have $|\langle \phi_T(\theta), p\rangle_T| \leq 1 - C_F$.

(iv) We have $\|p\|_T \leq C_B \sqrt{s}$.

The function $\eta : \theta \mapsto \langle \phi_T(\theta), p\rangle_T$ is the so-called "interpolating certificate" of the function v, as thanks to *(ii)* with $\theta = \theta^\star$, the function η coincides with the function v on $\mathcal{Q}^\star$. In addition, the interpolating certificate is required to have curvature properties in the near region and to be bounded by a constant strictly inferior to 1 in the far region. When r is sufficiently small (that is, $r \leq \sqrt{2/(C_N + C'_N)}$) Conditions *(i)* and *(ii)* are equivalent to the fact that the function η is in-between two quadratic functions in the near region of $\mathcal{Q}^\star$: for all $\theta^\star \in \mathcal{Q}^\star$ such that $v(\theta^\star) = 1$ (resp. $v(\theta^\star) = -1$) and $\theta \in \mathcal{B}_T(\theta^\star, r)$, we have $1 - C'_N\, \mathfrak{d}_T(\theta^\star, \theta)^2 \leq \eta(\theta) \leq 1 - C_N\, \mathfrak{d}_T(\theta^\star, \theta)^2$ (resp. $-1 + C_N\, \mathfrak{d}_T(\theta^\star, \theta)^2 \leq \eta(\theta) \leq -1 + C'_N\, \mathfrak{d}_T(\theta^\star, \theta)^2$).

Sufficient conditions for the next assumption to hold are also given in Section 2.7.

Assumption 2.6.2 (Interpolating derivative certificate)**.** *Let $T \in \mathbb{N}$, $s \in \mathbb{N}^*$, $r > 0$ and $\mathcal{Q}^\star$ be a subset of Θ_T of cardinal s. Suppose Assumptions 2.3.1 and 2.3.2 on the dictionary $(\varphi_T(\theta), \theta \in \Theta)$, and Assumption 2.5.1 on the kernel $\mathcal{K}_\infty$, defined on Θ^2, hold. Assume that $\mathfrak{d}_T(\theta, \theta') > 2r$ for all $\theta, \theta' \in \mathcal{Q}^\star \subset \Theta_T$ and that there exist finite positive constants c_N, c_F, c_B depending on r and $\mathcal{K}_\infty$, such that for any application $v : \mathcal{Q}^\star \to \{-1, 1\}$ there exists an element $q \in H_T$ satisfying:*

(i) For all $\theta^\star \in \mathcal{Q}^\star$ and $\theta \in \mathcal{B}_T(\theta^\star, r)$, we have:

$$|\langle \phi_T(\theta), q\rangle_T - v(\theta^\star)\ \mathrm{sign}(\theta - \theta^\star)\mathfrak{d}_T(\theta, \theta^\star)| \leq c_N\, \mathfrak{d}_T(\theta^\star, \theta)^2.$$

(ii) For all θ in Θ_T and $\theta \notin \bigcup_{\theta^\star \in \mathcal{Q}^\star} \mathcal{B}_T(\theta^\star, r)$ (far region), we have $|\langle \phi_T(\theta), q\rangle_T| \leq c_F$.

(iii) $||q||_T \leq c_B \sqrt{s}$.

The function $\theta \mapsto \langle \phi_T(\theta), q\rangle_T$ will be called an "interpolating derivative certificate" as it vanishes on $\mathcal{Q}^\star$. In addition, this function is required to decrease similarly to the function $\mathfrak{d}_T(\cdot, \theta^\star)$ near $\theta^\star$ and to be bounded in the far region of $\mathcal{Q}^\star$.

2.7 Sufficient conditions for the existence of certificates

In this section, we prove the existence of the certificate functions of Assumptions 2.6.1 and 2.6.2 provided that the parameters to be estimated are sufficiently separated in terms of the Riemannian metric. According to [Tang, 2015], the separation condition cannot be avoided to build certificate functions in general. It is however possible to remove this separation condition in some particular cases, see [Schiebinger et al., 2018] for models with positive amplitudes.

In order to find sufficient conditions for the existence of the interpolating certificate functions of Assumption 2.6.1, we extend the construction from [Poon et al., 2021] to a non asymptotic setting. For the existence of the interpolating derivative certificate functions of Assumption 2.6.2, we generalize the proof of [Candès and Fernandez-Granda, 2013, Lemma 2.7] dedicated to the dictionary of complex exponential functions. The proofs for the existence of certificates given in Section 2.10 require boundedness and local concavity properties of the kernel $\mathcal{K}_T$. For practical application, they are deduced from the boundedness and local concavity properties of the kernel $\mathcal{K}_\infty$ and the quality of approximation of $\mathcal{K}_T$ by $\mathcal{K}_\infty$ discussed in Section 2.5.

2.7.1 Boundedness and local concavity of the kernel $\mathcal{K}_T$

In this work, we shall consider bounded kernels locally concave on the diagonal. More precisely, for $T \in \bar{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$ and $r > 0$, we define:

$$\varepsilon_T(r) = 1 - \sup\left\{|\mathcal{K}_T(\theta, \theta')|; \quad \theta, \theta' \in \Theta_T \text{ such that } \mathfrak{d}_T(\theta', \theta) \geq r\right\}, \tag{2.35}$$

$$\nu_T(r) = -\sup\left\{\mathcal{K}_T^{[0,2]}(\theta, \theta'); \quad \theta, \theta' \in \Theta_T \text{ such that } \mathfrak{d}_T(\theta', \theta) \leq r\right\}. \tag{2.36}$$

The fact that $\varepsilon_T(r)$ and $\nu_T(r)$ are positive depends on the function φ_T, the space H_T and the set Θ_T. Let us mention that in many examples the positiveness of $\varepsilon_\infty(r)$ and $\nu_\infty(r)$ is easy to check whereas the positiveness of $\varepsilon_T(r)$ and $\nu_T(r)$ might be more difficult to prove.

Notice that (2.29) for $T \in \mathbb{N}$ and Assumption 2.5.1 for $T = \infty$, and the continuity of $\mathcal{K}_T$ and $\mathcal{K}_T^{[0,2]}$ give that:

$$\lim_{r \to 0^+} \varepsilon_T(r) = 0 \quad \text{and} \quad \lim_{r \to 0^+} \nu_T(r) = 1. \tag{2.37}$$

Recall ρ_T and $\mathcal{V}_T$ defined in (2.32) and (2.34). The next lemmas state that if $\varepsilon_\infty(r/\rho_T)$ (resp. $\nu_\infty(r\rho_T)$) is positive and if the approximation of $\mathcal{K}_T$ by $\mathcal{K}_\infty$ is good, *i.e.* $\mathcal{V}_T$ is small, then $\varepsilon_T(r)$ (resp. $\nu_T(r)$) is also positive.

Lemma 2.7.1. *Let $T \in \mathbb{N}$. Suppose Assumptions 2.3.1, 2.3.2 and 2.5.1 hold. Then we have for $r > 0$:*

$$\varepsilon_T(r) \geq \varepsilon_\infty(r/\rho_T) - \mathcal{V}_T \quad \text{and} \quad \nu_T(r) \geq \nu_\infty(r\rho_T) - \mathcal{V}_T.$$

Proof. As Assumptions 2.3.2 and 2.5.1 hold, recall that $\mathfrak{d}_\infty/\rho_T \leq \mathfrak{d}_T \leq \rho_T \mathfrak{d}_\infty$ on Θ_T^2, see (2.33).

Let $\theta, \theta' \in \Theta_T$ such that $\mathfrak{d}_T(\theta', \theta) \geq r$. We have $\mathfrak{d}_\infty(\theta', \theta) \geq r/\rho_T$. We get from the definition of $\mathcal{V}_T$ that:

$$|\mathcal{K}_T(\theta, \theta')| \leq |\mathcal{K}_\infty(\theta, \theta')| + \mathcal{V}_T \leq 1 - \varepsilon_\infty(r/\rho_T) + \mathcal{V}_T.$$

Then, use (2.35) to get $\varepsilon_T(r) \geq \varepsilon_\infty(r/\rho_T) - \mathcal{V}_T$. We also have $\mathfrak{d}_\infty(\theta', \theta) \leq r\rho_T$. We deduce that:

$$-\mathcal{K}_T^{[0,2]}(\theta, \theta') \geq -\mathcal{K}_\infty^{[0,2]}(\theta, \theta') - \mathcal{V}_T \geq \nu_\infty(r\rho_T) - \mathcal{V}_T.$$

Finally, using (2.36), we obtain $\nu_T(r) \geq \nu_\infty(r\rho_T) - \mathcal{V}_T$. □

When we require in addition of the assumptions of Lemma 2.7.1 that

$$\varepsilon_\infty(r/\rho_T) \wedge \nu_\infty(r\rho_T) > \mathcal{V}_T \geq 0,$$

then we have $\varepsilon_T(r) > 0$ and $\nu_T(r) > 0$.

2.7.2 Separation conditions for the non-linear parameters

In what follows, we measure the interferences (or the overlap) between the features in the mixture through a quantity δ_T introduced in [Poon et al., 2021] and defined below. Let $T \in \bar{\mathbb{N}}$, $\delta > 0$ and $s \in \mathbb{N}^*$. We define the set $\Theta^s_{T,\delta} \subset \Theta^s_T$ of vector of parameters of dimension $s \in \mathbb{N}^*$ and separation $\delta > 0$ as:

$$\Theta^s_{T,\delta} = \Big\{(\theta_1, \cdots, \theta_s) \in \Theta^s_T : \ \mathfrak{d}_T(\theta_\ell, \theta_k) > \delta \text{ for all distinct } k, \ell \in \{1, \dots, s\}\Big\}.$$

Using the convention $\inf \emptyset = +\infty$, we set for $u > 0$:

$$\delta_T(u, s) = \inf \Big\{\delta > 0 : \ \max_{1\leq \ell \leq s} \sum_{k=1, k\neq \ell}^{s} |\mathcal{K}_T^{[i,j]}(\theta_\ell, \theta_k)| \leq u \\ \text{for all } (i,j) \in \{0,1\} \times \{0,1,2\} \text{ and } (\theta_1, \cdots, \theta_s) \in \Theta^s_{T,\delta}\Big\}. \quad (2.38)$$

The quantity $\delta_T(u, s)$ is the minimum distance (with respect to the Riemannian metric $\mathfrak{d}_T$) between s parameters so that the coherence of the associated dictionary is bounded by u. The notion of coherence between the features in the definition of $\delta_T(u, s)$ is quite similar to the one used in compressed sensing (see [Foucart and Rauhut, 2013, Section 5]). A standard problem in compressed sensing is to retrieve the vector $\beta^\star$ when the multivariate function $\Phi_T(\vartheta^\star)$ is known in the discrete setting of Example 2.1.1 or Section 2.1.2.1. In this framework, the matrix $\Phi_T(\vartheta^\star)$, whose rows correspond to the K discretized functions in the dictionary, is known. The coherence is defined as $\max_{1\leq k\neq \ell \leq K} |\mathcal{K}_T(\theta^\star_k, \theta^\star_\ell)|$. Usually, the smaller the coherence, the easier it is to retrieve the parameter $\beta^\star$. The Babel function, introduced in [Tropp, 2004], is even closer to our measure of overlap. We refer to [Poon et al., 2021] for a discussion on this function.

Remark 2.7.2 (Rewriting the separation condition with operator norm). We shall stress that the definition of δ_T in (2.38) is related to the operator norm $\|\cdot\|_{\text{op}}$ associated to the ℓ_∞ norm on $\mathbb{R}^s$. We restate (2.38) using this operator norm $\|\cdot\|_{\text{op}}$, and leave the interested reader to check that another choice of operator norm does not improve the bounds on the certificates. Let us define for $i, j = 0, 1, 2$ (assuming the kernel $\mathcal{K}_T$ is smooth enough) and $\vartheta = (\theta_1, \dots, \theta_s) \in \Theta^s_T$ the $s \times s$ matrix:

$$\mathcal{K}_T^{[i,j]}(\vartheta) = \Big(\mathcal{K}_T^{[i,j]}(\theta_k, \theta_\ell)\Big)_{1\leq k,\ell\leq s}.$$

Let I be the identity matrix of size $s \times s$. For $i = 0$ or $i = 1$, since the diagonal coefficients of $\mathcal{K}_T^{[i,i]}(\vartheta)$ are equal to 1, see (2.25), we get:

$$\Big\|I - \mathcal{K}_T^{[i,i]}(\vartheta)\Big\|_{\text{op}} = \max_{1\leq k\leq s} \sum_{\ell\neq k} |\mathcal{K}_T^{[i,i]}(\theta_k, \theta_\ell)|.$$

Since the diagonal coefficients of $\mathcal{K}_T^{[1,0]}(\vartheta)$, $\mathcal{K}_T^{[0,1]}(\vartheta)$ and $\mathcal{K}_T^{[1,2]}(\vartheta)$ are zero, see (2.29), we also get:

$$\Big\|\mathcal{K}_T^{[1,0]}(\vartheta)\Big\|_{\text{op}} = \max_{1\leq k\leq s} \sum_{\ell\neq k} |\mathcal{K}_T^{[1,0]}(\theta_k, \theta_\ell)| \quad \text{and} \quad \Big\|\mathcal{K}_T^{[1,2]}(\vartheta)\Big\|_{\text{op}} = \max_{1\leq k\leq s} \sum_{\ell\neq k} |\mathcal{K}_T^{[1,2]}(\theta_k, \theta_\ell)|$$

and by symmetry, with $\|\cdot\|_{\mathrm{op}}^*$ for the operator norm associated to the ℓ_1 norm:

$$\left\|\mathcal{K}_T^{[0,1]}(\vartheta)\right\|_{\mathrm{op}} = \left\|\mathcal{K}_T^{[1,0]\top}(\vartheta)\right\|_{\mathrm{op}} = \left\|\mathcal{K}_T^{[1,0]}(\vartheta)\right\|_{\mathrm{op}}^* = \max_{1\le\ell\le s}\sum_{k\neq\ell}|\mathcal{K}_T^{[1,0]}(\theta_k,\theta_\ell)|$$
$$= \max_{1\le k\le s}\sum_{\ell\neq k}|\mathcal{K}_T^{[0,1]}(\theta_k,\theta_\ell)|.$$

Since the diagonal coefficients of $\mathcal{K}_T^{[2,0]}(\vartheta)$ are equal to -1, see (2.29), we also get:

$$\left\|I+\mathcal{K}_T^{[2,0]}(\vartheta)\right\|_{\mathrm{op}} = \max_{1\le k\le s}\sum_{\ell\neq k}|\mathcal{K}_T^{[2,0]}(\theta_k,\theta_\ell)|.$$

Thus, we have:

$$\delta_T(u,s) = \inf\left\{\delta>0:\ A_{T,\ell_\infty}(\vartheta)\le u, \vartheta\in\Theta^s_{T,\delta}\right\}, \tag{2.39}$$

where:

$$A_{T,\ell_\infty}(\vartheta) = \max\Big(\left\|I-\mathcal{K}_T^{[0,0]}(\vartheta)\right\|_{\mathrm{op}}, \left\|I-\mathcal{K}_T^{[1,1]}(\vartheta)\right\|_{\mathrm{op}}, \left\|I+\mathcal{K}_T^{[2,0]}(\vartheta)\right\|_{\mathrm{op}},$$
$$\left\|\mathcal{K}_T^{[1,0]}(\vartheta)\right\|_{\mathrm{op}}, \left\|\mathcal{K}_T^{[0,1]}(\vartheta)\right\|_{\mathrm{op}}, \left\|\mathcal{K}_T^{[1,2]}(\vartheta)\right\|_{\mathrm{op}}\Big). \tag{2.40}$$

Lemma 2.7.3 below enables us to compare the separation distance at T fixed and at the limit case where $T=+\infty$. Recall that the constant ρ_T is defined in (2.32).

Lemma 2.7.3. *Let $T\in\bar{\mathbb{N}}$ and $s\in\mathbb{N}^*$. Suppose Assumptions 2.3.1, 2.3.2 and 2.5.1 hold. Then, for $u>0$ and with:*

$$u_T(s) = u+(s-1)\mathcal{V}_T,$$

we have:

$$\delta_T(u_T(s),s)\le\rho_T\,\delta_\infty(u,s) \quad \text{and} \quad \Theta^s_{T,\rho_T\,\delta_\infty(u,s)}\subseteq\Theta^s_{T,\delta_T(u_T(s),s)}.$$

Proof. Since Assumptions 2.3.2 and 2.5.1 hold, we have from (2.33) that $\mathfrak{d}_T\le\rho_T\,\mathfrak{d}_\infty$ on Θ_T^2. Hence for any $\delta>0$, we have the inclusion $\Theta^s_{T,\rho_T\,\delta}\subseteq\Theta^s_{\infty,\delta}$. In particular, we have for $u>0$ that $\Theta^s_{T,\rho_T\,\delta_\infty(u,s)}\subseteq\Theta^s_{\infty,\delta_\infty(u,s)}$. Using the triangle inequality and the definition of $\mathcal{V}_T$ in (2.34), we have that for $(i,j)\in\{0,1\}\times\{0,1,2\}$ and $(\theta_1,\cdots,\theta_s)\in\Theta_T^s$:

$$\sum_{k=1,k\neq\ell}^{s}|\mathcal{K}_T^{[i,j]}(\theta_\ell,\theta_k)|\le\sum_{k=1,k\neq\ell}^{s}\left(|\mathcal{K}_\infty^{[i,j]}(\theta_\ell,\theta_k)|+\mathcal{V}_T\right).$$

Then, the inclusion $\Theta^s_{T,\rho_T\,\delta_\infty(u,s)}\subseteq\Theta^s_{\infty,\delta_\infty(u,s)}$ gives that for all $(i,j)\in\{0,1\}\times\{0,1,2\}$ and $(\theta_1,\cdots,\theta_s)\in\Theta^s_{T,\rho_T\,\delta_\infty(u,s)}$:

$$\sum_{k=1,k\neq\ell}^{s}|\mathcal{K}_T^{[i,j]}(\theta_\ell,\theta_k)|\le u+(s-1)\mathcal{V}_T.$$

With $u_T(s)=u+(s-1)\mathcal{V}_T$, we deduce that $\delta_T(u_T(s),s)\le\rho_T\,\delta_\infty(u,s)$, which proves the inclusion $\Theta^s_{T,\rho_T\,\delta_\infty(u,s)}\subseteq\Theta^s_{T,\delta_T(u_T(s),s)}$. □

2.7.3 The interpolating certificates

We define quantities which depend on $\mathcal{K}_\infty$, Θ_∞ and on real parameters $r>0$ and $\rho\ge1$:

$$\begin{aligned}H_\infty^{(1)}(r,\rho) &= \frac{1}{2}\wedge L_{2,0}\wedge L_{2,1}\wedge\frac{\nu_\infty(\rho r)}{10}\wedge\frac{\varepsilon_\infty(r/\rho)}{10},\\ H_\infty^{(2)}(r,\rho) &= \frac{1}{6}\wedge\frac{8\,\varepsilon_\infty(r/\rho)}{10(5+2L_{1,0})}\wedge\frac{8\,\nu_\infty(\rho r)}{9(2L_{2,0}+2L_{2,1}+4)},\end{aligned} \tag{2.41}$$

where the constants involved are defined in (2.31). By recalling the behaviors of $\varepsilon_\infty(r)$ and $\nu_\infty(r)$ when r goes down to zero from (2.37), we have for $\rho \geq 1$:

$$\lim_{r\to 0^+} H^{(1)}_\infty(r,\rho) = 0 \quad \text{and} \quad \lim_{r\to 0^+} H^{(2)}_\infty(r,\rho) = 0.$$

We state the first main result of this section whose proof is given in Section 2.10.

Proposition 2.7.4 (Interpolating certificate)**.** *Let $T \in \mathbb{N}$, $s \in \mathbb{N}^*$, $\rho \geq 1$ and $r > 0$. We assume that:*

(i) ***Regularity of the dictionary φ_T:*** *Assumptions 2.3.1 and 2.3.2 hold.*

(ii) ***Regularity of the limit kernel $\mathcal{K}_\infty$:*** *Assumption 2.5.1 holds. Furthermore, we have $r \in (0, 1/\sqrt{2L_{2,0}})$, and also $\varepsilon_\infty(r/\rho) > 0$ and $\nu_\infty(\rho r) > 0$.*

(iii) ***Separation of the non-linear parameters:*** *There exists $u_\infty \in \left(0, H^{(2)}_\infty(r,\rho)\right)$ such that:*

$$\delta_\infty(u_\infty, s) < +\infty.$$

(iv) ***Closeness of the metrics $\mathfrak{d}_T$ and $\mathfrak{d}_\infty$:*** *We have $\rho_T \leq \rho$.*

(v) ***Proximity of the kernels $\mathcal{K}_T$ and $\mathcal{K}_\infty$:***

$$\mathcal{V}_T \leq H^{(1)}_\infty(r,\rho) \quad \text{and} \quad (s-1)\mathcal{V}_T \leq H^{(2)}_\infty(r,\rho) - u_\infty.$$

Then, with the positive constants:

$$C_N = \frac{\nu_\infty(\rho r)}{180}, \quad C'_N = \frac{5}{8}L_{2,0} + \frac{1}{8}L_{2,1} + \frac{1}{2}, \quad C_B = 2 \quad \text{and} \quad C_F = \frac{\varepsilon_\infty(r/\rho)}{10} \leq 1, \tag{2.42}$$

Assumption 2.6.1 holds (with the same r) for any subset $\mathcal{Q}^\star = \{\theta^\star_i,\, 1 \leq i \leq s\}$ such that for all $\theta \neq \theta' \in \mathcal{Q}^\star$:

$$\mathfrak{d}_T(\theta, \theta') > 2\,\max(r, \rho_T\,\delta_\infty(u_\infty, s)).$$

Note that *(i)* concerns the dictionary φ_T, *(ii)* and *(iii)* the limit kernel $\mathcal{K}_\infty$ and the set of parameters, and *(iv)* and *(v)* the regime for the parameters s and T. Notice that if $\mathcal{K}_\infty$ is chosen equal to $\mathcal{K}_T$, then $\mathcal{V}_T = 0$ and $\rho_T = 1$, and also *(iv)* and *(v)* hold and ρ can be chosen equal to 1.

We now give the second main result of this section whose proof is given in Section 2.10.2.

Proposition 2.7.5 (Interpolating derivative certificate)**.** *Let $T \in \mathbb{N}$ and $s \in \mathbb{N}^*$. We assume that:*

(i) ***Regularity of the dictionary φ_T:*** *Assumptions 2.3.1 and 2.3.2 hold.*

(ii) ***Regularity of the limit kernel $\mathcal{K}_\infty$:*** *Assumption 2.5.1 holds.*

(iii) ***Separation of the non-linear parameters:*** *There exists $u'_\infty \in (0, 1/6)$, such that:*

$$\delta_\infty(u'_\infty, s) < +\infty.$$

(iv) ***Proximity of the kernels $\mathcal{K}_T$ and $\mathcal{K}_\infty$:*** *We have:*

$$\mathcal{V}_T \leq 1 \quad \text{and} \quad (s-1)\mathcal{V}_T + u'_\infty \leq 1/6.$$

Then, with the positive constants:

$$c_N = \frac{1}{8}L_{2,0} + \frac{5}{8}L_{2,1} + \frac{7}{8}, \quad c_B = 2 \quad \textit{and} \quad c_F = \frac{5}{4}L_{1,0} + \frac{7}{4}, \tag{2.43}$$

Assumption 2.6.2 holds for any $r > 0$ and any subset $\mathcal{Q}^\star = \{\theta_i^\star,\, 1 \le i \le s\}$ such that for all $\theta \neq \theta' \in \mathcal{Q}^\star$:

$$\mathfrak{d}_T(\theta, \theta') > 2\, \max(r, \rho_T\, \delta_\infty(u'_\infty, s)).$$

Let us briefly indicate how the certificates are constructed in Section 2.10 using the features of the dictionary. Let $\alpha = (\alpha_1, \dots, \alpha_s)$ and $\xi = (\xi_1, \dots, \xi_s)$ be elements of $\mathbb{R}^s$. Let $p_{\alpha,\xi} \in H_T$ be defined by:

$$p_{\alpha,\xi} = \sum_{k=1}^{s} \alpha_k \phi_T(\theta_k^\star) + \sum_{k=1}^{s} \xi_k\, \phi_T^{[1]}(\theta_k^\star),$$

where $\phi_T^{[1]}$ denotes the derivative $\tilde{D}_{1;T}[\phi_T]$. Using (2.28) in Lemma 2.4.3, set the interpolating real-valued function $\eta_{\alpha,\xi}$ defined on Θ by:

$$\eta_{\alpha,\xi}(\theta) := \langle \phi_T(\theta), p_{\alpha,\xi} \rangle_T = \sum_{k=1}^{s} \alpha_k\, \mathcal{K}_T(\theta, \theta_k^\star) + \sum_{k=1}^{s} \xi_k\, \mathcal{K}_T^{[0,1]}(\theta, \theta_k^\star).$$

By Assumption 2.3.2 on the regularity of φ_T and the positivity of g_T and Lemma 2.4.3, we get that the function $\eta_{\alpha,\xi}$ is of class $\mathcal{C}^3$ on Θ, and using (2.20), we get that:

$$\eta_{\alpha,\xi}^{[1]} := \tilde{D}_{1;T}[\eta_{\alpha,\xi}](\theta) = \sum_{k=1}^{s} \alpha_k\, \mathcal{K}_T^{[1,0]}(\theta, \theta_k^\star) + \sum_{k=1}^{s} \xi_k\, \mathcal{K}_T^{[1,1]}(\theta, \theta_k^\star).$$

We show in Section 2.10 that for any function $v : \mathcal{Q}^\star \to \{-1, 1\}$ there exists a unique choice of α and ξ such that $\eta_{\alpha,\xi}$ becomes an interpolating certificate, that is, $\eta_{\alpha,\xi} = v$ and $\eta_{\alpha,\xi}^{[1]} = 0$ on $\mathcal{Q}^\star$, and $p_{\alpha,\xi}$ satisfies Points *(i)*-*(iv)* of Assumption 2.6.1.

Moreover, for any function $v : \mathcal{Q}^\star \to \{-1, 1\}$ there exists another unique choice of α and ξ such that $\eta_{\alpha,\xi}$ is an interpolating derivative certificate, that is, $\eta_{\alpha,\xi} = 0$ and $\eta_{\alpha,\xi}^{[1]} = v$ on $\mathcal{Q}^\star$, and $p_{\alpha,\xi}$ satisfies Points *(i)*-*(iii)* of Assumption 2.6.2.

2.8 Sparse spike deconvolution

We develop here in full details the particular example of a mixture of Gaussian features observed in a discrete regression model with regular design. In particular, we check the numerous but not very restrictive assumptions, and we illustrate that our general and more restrictive sufficient conditions for the existence of certificates can turn simpler and far less restrictive on concrete examples. The model is presented in Section 2.8.1, where we also check the first assumptions. The technical Section 2.8.2 on the existence of the certificates allows to point out the separation distance in (2.50) and with the simpler expression in (2.51). This separation distance is usually very pessimistic, but one can rely on numerical estimations to be more realistic, see Remark 2.8.2 in this direction. Eventually, we apply to this context our main Theorem 2.2.1 in Section 2.8.3 as Corollary 2.8.3 and illustrate a particular choice of the tuning parameter in Remark 2.8.4 in the spirit of [Tang et al., 2015, Boyer et al., 2017] established for the specific dictionary of complex exponentials.

2.8.1 Model and first assumptions of Theorem 2.2.1

Consider a real-valued process y observed over a regular grid $t_1 < \cdots < t_T$ of a symmetric interval $[a_T, b_T]$, with $T \ge 2$, $a_T = -b_T < 0$, $t_j = a_T + j\Delta_T$ for $j = 1, \dots, T$ and grid step:

$$\Delta_T = \frac{b_T - a_T}{T}.$$

Assuming that all the observations have the same weight amounts to considering y as an element of the Hilbert space $H_T = L^2(\lambda_T)$ of real valued functions defined in $\mathbb{R}$ and square integrable with respect to the atomic measure λ_T on $\{t_1, \dots, t_T\}$:

$$\lambda_T(\mathrm{d}t) = \Delta_T \sum_{j=1}^{T} \delta_{t_j}(\mathrm{d}t).$$

We consider a noise process $w_T(t) = \sum_{j=1}^{T} G_j \mathbf{1}_{\{t_j = t\}}$ for $t \in \mathbb{R}$, where $(G_1, \dots, G_T)$ is a centered Gaussian vector such that, for some noise level $\sigma_1 > 0$:

$$\mathbb{E}[G_j^2] = \sigma_1^2 \quad \text{and} \quad |\mathbb{E}[G_j G_i]| \leq \sigma_1^2/T \quad \text{for } j \neq i \text{ in } \{1,\dots,\mathrm{T}\}.$$

Thus, the norm of the noise $\|w_T\|_T$ is finite almost surely, and for any $f \in L^2(\lambda_T)$ we have:

$$\mathrm{Var}(\langle f, w_T \rangle_T) = \mathrm{Var}\Big(\Delta_T \sum_{j=1}^{T} f(t_j) G_j\Big) \leq 2\sigma_1^2 \Delta_T \|f\|_T^2.$$

Hence, Assumption 2.1.1 on the noise is satisfied with $\sigma^2 = 2\sigma_1^2$. (Notice that if the random variables $G_1, \dots, G_T$ are independent, then $\mathrm{Var}(\langle f, w_T \rangle_T) = \sigma^2 \Delta_T \|f\|_T^2$ with $\sigma^2 = \sigma_1^2$.) This gives that Point *(i)* of Theorem 2.2.1 holds.

We consider the dictionary given by the translation model of Section 2.3.3.1 with Gaussian features and fixed scaling parameter $\sigma_0 > 0$, that is the dictionary does not depend on T and is given by:

$$\Big(\varphi(\theta) = k\Big(\frac{\cdot - \theta}{\sigma_0}\Big),\, \theta \in \Theta\Big) \quad \text{with} \quad k(t) = \mathrm{e}^{-t^2/2} \quad \text{and} \quad \Theta = \mathbb{R}.$$

Thus, the signal $\beta^\star \Phi(\vartheta^\star)$ in model (2.2) can indeed be written as the convolution product of the function k and an atomic measure. It is elementary to check that Assumption 2.3.1 on the regularity of the features holds. Furthermore, the functions $\varphi(\theta)$ and $\partial_\theta \varphi(\theta)$ are linearly independent $\lambda_T - a.e$ for all $\theta \in \Theta$ as $T \geq 2$. Hence the function g_T is positive on Θ by Lemma 2.3.1 and thus Assumption 2.3.2 holds. This gives that Point *(ii)* of Theorem 2.2.1 holds.

We now define the limit kernel $\mathcal{K}_\infty$. To do so, we shall assume that $(b_T, T \geq 2)$ is a sequence of positive numbers, such that:

$$\lim_{T \to \infty} b_T = +\infty \quad \text{and} \quad \lim_{T \to \infty} \Delta_T = 0. \tag{2.44}$$

This in particular implies that the sequence of measures $(\lambda_T, T \geq 2)$ converges with respect to the vague topology towards the Lebesgue measure, say λ_∞, on $\Theta_\infty = \mathbb{R}$. We also consider the Hilbert space $H_\infty = L^2(\lambda_\infty)$ endowed with its usual scalar product denoted $\langle \cdot, \cdot \rangle_\infty$ and corresponding norm denoted $\|\cdot\|_\infty$ (not to be confused with the supremum norm!). Note that the kernel $\mathcal{K}_T$ and the associated quantities such as ε_T and ν_T defined in (2.35) and (2.36), respectively, or the uniform bounds on $\mathcal{K}_T^{[i,j]}$, are difficult to calculate. However the uniform bounds on $\Theta_\infty = \mathbb{R}$ for the kernel $\mathcal{K}_\infty$, defined by (2.26) with T replaced by ∞, are easily computed. Elementary calculations give for $\theta, \theta' \in \Theta$:

$$\|\varphi(\theta)\|_\infty^2 = \sqrt{\pi}\,\sigma_0, \quad \phi_\infty(\theta) = \frac{1}{\pi^{\frac{1}{4}} \sqrt{\sigma_0}} \varphi(\theta), \quad \mathcal{K}_\infty(\theta, \theta') = k\Big(\frac{\theta - \theta'}{\sqrt{2}\,\sigma_0}\Big) \quad \text{and} \quad g_\infty(\theta) = \frac{1}{2\sigma_0^2}.$$

In particular, we have $g'_\infty(\theta) = 0$. The Riemannian metric is equal to the Euclidean distance up to a multiplicative factor, for all $\theta, \theta' \in \Theta_\infty = \mathbb{R}$:

$$\mathfrak{d}_\infty(\theta, \theta') = \frac{|\theta - \theta'|}{\sqrt{2}\,\sigma_0}. \tag{2.45}$$

We see that $\mathcal{K}_\infty$ is of class $\mathcal{C}^{\infty,\infty}$ and that:

$$\mathcal{K}_\infty^{[i,j]}(\theta,\theta') = (-1)^j\, k^{(i+j)}\Big(\frac{\theta-\theta'}{\sqrt{2}\,\sigma_0}\Big) \quad \text{and} \quad k^{(i)}(t) = P_i(t)\,k(t), \tag{2.46}$$

where we give for convenience the formulae for some of the polynomials P_i:

$$P_1(t) = -t, \quad P_2(t) = -1+t^2, \quad P_3(t) = 3t - t^3,$$
$$P_4(t) = 3 - 6t^2 + t^4, \quad P_6(t) = -15 + 45t^2 - 15t^4 + t^6.$$

Then, we explicitly compute the constants $L_{i,j}$ for $i,j \in \{0,\cdots,2\}$ and L_3 defined in (2.31):

$$m_g = (2\sigma_0^2)^{-1}, \quad L_{0,0} = 1, \quad L_{1,0} = L_{0,1} = \mathrm{e}^{-1/2}, \quad L_{1,1} = L_{2,0} = L_{0,2} = 1,$$
$$L_{2,1} = L_{1,2} = \sqrt{18-6\sqrt{6}}\,\mathrm{e}^{\sqrt{3/2}-3/2} \le \sqrt{2}, \quad L_{2,2} = 3 \quad \text{and} \quad L_3 = 15.$$

Notice the constants $L_{i,j}$ and L_3 do not depend on the scaling factor σ_0. Thus Assumption 2.5.1 holds. This gives that Point (iii) of Theorem 2.2.1 holds.

We now check the proximity of the kernel $\mathcal{K}_T$ to the limit kernel $\mathcal{K}_\infty$. The support of λ_T is spread over the window $[a_T, b_T]$ where the signal is observed. Hence it is legitimate to look for the location parameters on a smaller subset of this window, and thus restrict the optimization (2.3) to the compact set:

$$\Theta_T = [(1-\epsilon)a_T, (1-\epsilon)b_T] \subset [a_T, b_T] \quad \text{with a given shrinkage } \epsilon > 0.$$

The proof of the next lemma is given in Section 2.11.4. Recall ρ_T and $\mathcal{V}_T$ defined in (2.32) and (2.34). Set:

$$\gamma_T = 2\Delta_T\,\sigma_0^{-1} + \sqrt{\pi}\;\mathrm{e}^{-\epsilon^2 b_T^2/2\sigma_0^2}\,.$$

Lemma 2.8.1. *There exist finite positive universal constants c_0, c_1 and c_2, such that $\gamma_T < c_0$ implies:*

$$\mathcal{V}_T \le c_1\gamma_T \quad \textit{and} \quad |1-\rho_T| \le c_2\gamma_T. \tag{2.47}$$

This implies that Assumption 2.5.2 holds for T such that $\gamma_T \le c_0$ and $c_1\gamma_T \le 3$, which holds for T large enough thanks to (2.44). Thus Point (iv) of Theorem 2.2.1 holds for T large enough.

2.8.2 Existence of certificates

We keep the model and the notations from Section 2.8.1. In order to get the prediction error from Theorem 2.2.1, we only need to check that Point (iv) therein on the existence of the certificates holds. To check the existence of the certificates, we can use Propositions 2.7.4 and 2.7.5, and check that all the hypotheses required in those two propositions hold.

We first concentrate on the hypotheses of Proposition 2.7.4. Assumption (i) on the regularity of the dictionary holds, see Section 2.8.1.

We recall that $L_{0,2} = 1$ and thus $1/\sqrt{2L_{0,2}} = 1/\sqrt{2} > 1/2$. Recall $\varepsilon_\infty(r)$ and $\nu_\infty(r)$ defined in (2.35) and (2.36), and thanks to the explicit form of the Riemannan metric, we get for $r \in (0,1)$:

$$\varepsilon_\infty(r) = 1 - \mathrm{e}^{-r^2/2} > 0 \quad \text{and} \quad \nu_\infty(r) = \left(1-r^2\right)\mathrm{e}^{-r^2/2}\,.$$

This and the regularity of the kernel $\mathcal{K}_\infty$ from Section 2.8.1 imply that Assumption (ii) holds for all $r \in (0, 1/(\rho \vee \sqrt{2}))$.

We obtain from (2.46) that $\lim_{q\to\infty} \sup_{|\theta-\theta'|\geq q} |\mathcal{K}_\infty^{[i,j]}(\theta,\theta')| = 0$ for all $i,j \in \{0,1,2\}$. Thus, we deduce from the definition (2.38) of δ_∞ that $\delta_\infty(u,s)$ is finite for all $s \in \mathbb{N}^*$ and $u > 0$. This implies that Assumption *(iii)* on the separation of the parameters holds.

To simplify, we set $\rho = 2$ (but we could take any value of $\rho > 1$). We deduce from Lemma 2.8.1, that for T large enough $\rho_T \leq \rho = 2$, and thus Assumption *(iv)* on the closeness of the metrics $\mathfrak{d}_T$ and $\mathfrak{d}_\infty$ holds.

Recall the definition of $H_\infty^{(1)}$ and $H_\infty^{(2)}$ from (2.41). To get the smallest separation distance, we also set:

$$r = \operatorname*{argmax}_{0\leq r'\leq 1/2} H_\infty^{(2)}(r',\rho) \approx 0.49. \tag{2.48}$$

Notice that the function is not *a priori* monotone in ρ. We have $\varepsilon_\infty(r/2) \approx 2.9\times 10^{-2}$, $\nu_\infty(2r) \approx 3.7\times 10^{-2}$, $H_\infty^{(1)}(r,2) \approx 2.9\times 10^{-3}$ and $H_\infty^{(2)}(r,2) \approx 3.7\times 10^{-3}$. Again in order to get a "small" separation distance, we choose u_∞ close to $H_\infty^{(2)}(r,2)$, say $u_\infty = \eta_0 H_\infty^{(2)}(r,2)$ for some $\eta_0 < 1$ close to 1. For simplicity set $\eta_0 = 9/10$. Thanks to hypothesis (2.44), we get $\lim_{T\to\infty}\gamma_T = 0$ and Lemma 2.8.1 implies that for T large enough, depending on σ_0, ϵ and the sparsity parameter s, we have:

$$\rho_T \leq 2, \quad \mathcal{V}_T \leq H_\infty^{(1)}(r,2) \quad \text{and} \quad (s-1)\mathcal{V}_T \leq (1-\eta_0)H_\infty^{(2)}(r,2), \tag{2.49}$$

and thus Assumption *(v)* on the proximity of the kernels $\mathcal{K}_T$ and $\mathcal{K}_\infty$ holds.

Thus, the assumptions of Proposition 2.7.4 are satisfied, and we deduce that Assumption 2.6.1 holds with, thanks to (2.42):

$$C_N \approx 2\times 10^{-4}, \quad C_N' \approx 1.3, \quad C_B = 2 \quad \text{and} \quad C_F \approx 2.9\times 10^{-3}.$$

We now concentrate on the hypotheses of Proposition 2.7.5. Assumptions *(i)*-*(iii)* clearly hold for the same reasons as Assumptions *(i)*-*(iii)* of Proposition 2.7.4.

Again in order to get a "small" separation distance, there is no need to choose u_∞' larger that u_∞, and for this reason we take $u_\infty' = u_\infty$. We deduce from (2.49) that for T large enough, depending on σ_0, ϵ and the sparsity parameter s:

$$\mathcal{V}_T \leq 1 \quad \text{and} \quad (s-1)\mathcal{V}_T + u_\infty' \leq 1/6,$$

and thus Assumption *(iv)* on the proximity of the kernels $\mathcal{K}_T$ and $\mathcal{K}_\infty$ holds.

Thus, the assumptions of Proposition 2.7.5 are satisfied, and we deduce, thanks to (2.43), that Assumption 2.6.2 holds with the same value of r given by (2.48):

$$c_N \approx 1.9, \quad c_B = 2, \quad \text{and} \quad c_F \approx 2.6.$$

In conclusion, we get that Assumptions 2.6.1 and 2.6.2 hold for T large enough, and thus Point *(v)* of Theorem 2.2.1 holds for T large enough and $\mathcal{Q}^\star$ such that for all $\theta \neq \theta' \in \mathcal{Q}^\star$ the distance $\mathfrak{d}_T(\theta,\theta')$ is larger than the separation distance:

$$2\max(r,\ \rho_T\,\delta_\infty(u_\infty,s),\ \rho_T\,\delta_\infty(u_\infty',s)). \tag{2.50}$$

Notice that since $u_\infty = u_\infty'$, $\rho_T\mathfrak{d}_T(\theta,\theta') \geq \mathfrak{d}_\infty(\theta,\theta')$ and $\rho_T \leq 2$, we deduce from (2.45), that a slightly stronger condition is to assume that $|\theta-\theta'|$ is larger than:

$$\sqrt{2}\,\sigma_0 \max(1, 4\delta_\infty(u_\infty,s)). \tag{2.51}$$

Remark 2.8.2 (On the separation distance (2.50)). The separation distance (2.50) is a non-decreasing function of s. We now provide an upper bound. Let $(i,j) \in \{0,1\} \times \{0,1,2\}$. By considering the kernel $\mathcal{K}_T$ and its derivative given by (2.46) and the bound

$$M = \max_{0 \leq i \leq 3} \sup |P_i| \sqrt{k},$$

we deduce that $|\mathcal{K}_\infty^{[i,j]}(\theta,\theta')| \leq M \, \mathrm{e}^{-\mathfrak{d}_\infty(\theta,\theta')^2/2}$ for all $\theta, \theta' \in \Theta$. We easily obtain that for $\vartheta = (\theta_1, \cdots, \theta_s) \in \Theta^s_{\infty,\delta}$ with $\delta > 0$:

$$\max_{1 \leq \ell \leq s} \sum_{k=1, k \neq \ell}^{s} |\mathcal{K}_\infty^{[i,j]}(\theta_\ell, \theta_k)| \leq \psi_s(\delta) \quad \text{with} \quad \psi_s(\delta) = 2M \int_0^{s/2+1} \mathrm{e}^{-t^2\delta^2/4} \, \mathrm{d}t.$$

The function ψ_s is decreasing and one to one from $\mathbb{R}_+$ to $(0, M(s+2)]$. Setting $\psi_s^{-1}(u) = 0$ for $u > M(s+2)$, we deduce from (2.38) that for $u > 0$:

$$\delta_\infty(u,s) \leq \psi_s^{-1}(u).$$

Since the map $s \mapsto \psi_s(\delta)$ is increasing with limit $\psi_\infty(\delta) = 2\sqrt{\pi}\, M/\delta$, we deduce that for $s \in \mathbb{N}^*$:

$$\delta_\infty(u,s) \leq \frac{2\sqrt{\pi}\, M}{u},$$

so that the separation distance (2.50) (or (2.51)) can be bounded uniformly in s for given r and $u_\infty = u'_\infty$.

In fact, we shall illustrate for $s = 2$ that the separation distance (2.50) is largely overestimated. We can compute $\delta_\infty(u,s)$ thanks to its expression (2.39). For $s = 2$ and with the values chosen in this section for $u_\infty = u'_\infty$, we obtain $\delta_\infty(u_\infty, 2) \approx 4.5$. We deduce that the separation distance (2.50) expressed with respect to the metric $\mathfrak{d}_T$ is approximately $9\,\rho_T$ (which gives $13\,\sigma_0\rho_T^2$ in terms of the Euclidean metric), which is unconveniently large. However, a detailed numerical approach (using the very certificates provided in the proof of Propositions 2.7.4 and 2.7.5) with T large so that the kernel $\mathcal{K}_T$ is indeed well approximated by $\mathcal{K}_\infty$ (and thus $\rho_T \approx 1$), gives that one can take for $s = 2$ the separation distance with respect to the Euclidean metric equal to $3.1 \times \sigma_0$ (that is approximately equal to 2.2 with respect to the metric $\mathfrak{d}_\infty$), which is much more realistic. Therefore, the theoretical separation distance (2.50) is in general largely overestimated. We represent in Figures 2.1 and 2.2 interpolating certificates and interpolating derivative certificates for a set of parameters $\mathcal{Q}^\star = \{\theta_1^\star, \theta_2^\star\}$ such that $|\theta_1^\star - \theta_2^\star| \geq 3.1 \times \sigma_0$. These certificates satisfy the interpolating, boundedness and curvature properties required in Assumption 2.6.1 and 2.6.2. We also observe that when the parameters in $\mathcal{Q}^\star$ are too close it may be impossible to obtain a certificate that satisfies Assumption 2.6.1 from the construction presented in Section 2.7.3, see Figure 2.3 where the boundeness condition is not met.

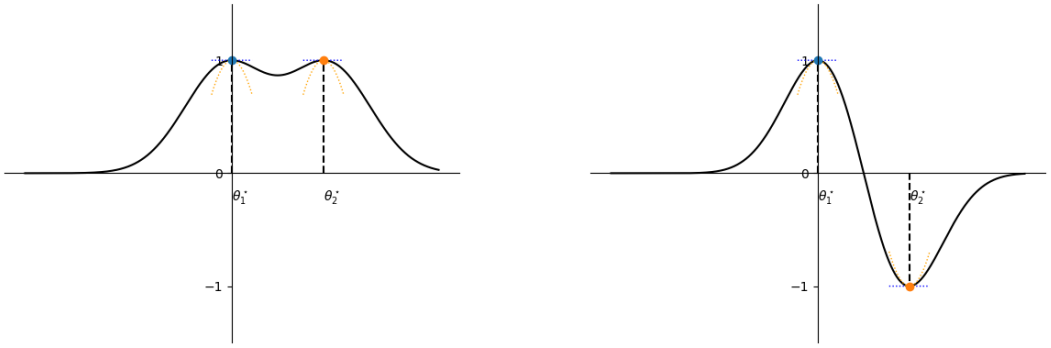

Figure 2.1 – Interpolating certificates satisfying the interpolating, boundedness and curvature properties from Assumption 2.6.1.

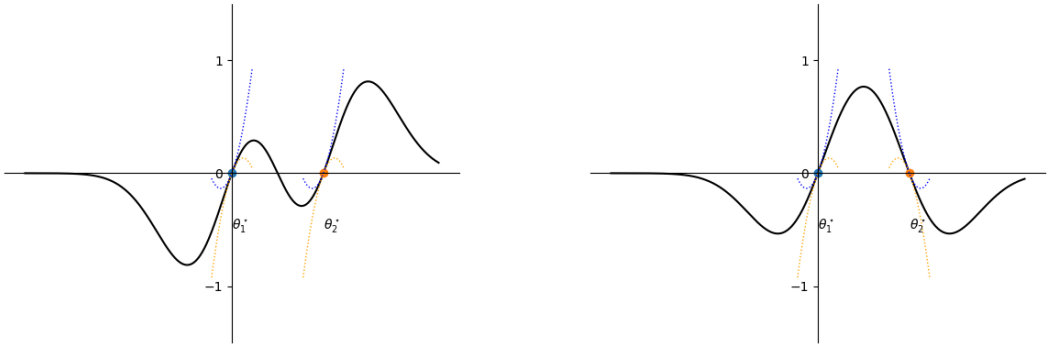

Figure 2.2 – Interpolating derivative certificates satisfying the interpolating, boundedness and curvature properties from Assumption 2.6.2.

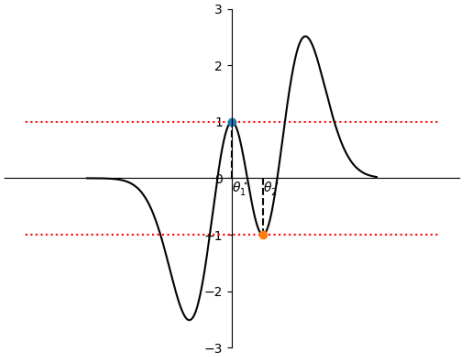

Figure 2.3 – Interpolating certificate violatiating the boundedness condition of Assumption 2.6.1 (*i.e* $\sup_\theta |\eta(\theta)| > 1 - C_F$).

2.8.3 Prediction error

We keep the model and the notations from Section 2.8.1 and the values chosen in Section 2.8.2. We deduce from Theorem 2.2.1 the following result.

Corollary 2.8.3. *For T large enough, depending on σ_0, ϵ and the sparsity parameter s, such*

that (2.49) holds and for all $\theta \neq \theta' \in \mathcal{Q}^\star = \{\theta_k^\star,\, k \in S^\star\}$, *with* $S^\star = \mathrm{Supp}(\beta^\star)$ *such that* $|\theta - \theta'|$ *is larger than the separation parameter* $\sqrt{2}\,\sigma_0 \max(1, 4\delta_\infty(u_\infty, s))$ *given by (2.51), then, with some universal finite constants* $\mathcal{C}_0, ..., \mathcal{C}_3 > 0$, *for any* $\tau > 1$ *and a tuning parameter:*

$$\kappa \geq \mathcal{C}_1 \sigma \sqrt{\Delta_T \log(\tau)}, \tag{2.52}$$

we have the prediction error bound of the estimators $\hat{\beta}$ *and* $\hat{\vartheta}$ *defined in (2.3) given by:*

$$\sqrt{\Delta_T}\left\|\hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star\Phi_T(\vartheta^\star)\right\|_{\ell_2} \leq \mathcal{C}_0\,\sqrt{s}\,\kappa,$$

with probability larger than $1 - \mathcal{C}_2\left(\frac{\sqrt{2}b_T}{\sigma_0\tau\sqrt{\log(\tau)}} \vee \frac{1}{\tau}\right)$. *Moreover, with the same probability, we have that* $\left|\|\hat{\beta}\|_{\ell_1} - \|\beta^\star\|_{\ell_1}\right| \leq \mathcal{C}_3\kappa s$ *as well as the inequalities (2.7) of Theorem 2.2.5.*

The values of the universal constants $\mathcal{C}_i$, $i = 0, \ldots, 3$, can be given explicitly and they are large, but they could be improved numerically.

Remark 2.8.4 (A particular choice of the tuning parameter). Let $\gamma > 0$ and $\gamma' \geq \gamma$ such that $1 > \gamma' - \gamma$. Set $\tau = T^{\gamma'}$, $b_T = \sigma_0 T^{\gamma'-\gamma}\sqrt{\log(T)}$ and $\kappa = \mathcal{C}_1\sigma\sqrt{\Delta_T \log(\tau)}$ (which corresponds to the equality in (2.52)). Then, we get under the assumptions of Corollary 2.8.3 (and thus T large enough) that:

$$\frac{1}{\sqrt{T}}\left\|\hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star\Phi_T(\vartheta^\star)\right\|_{\ell_2} \leq \mathcal{C}_0''\,\sigma\sqrt{s\frac{\log T}{T}},$$

with probability larger than $1 - \mathcal{C}_2''/T^\gamma$ where $\mathcal{C}_0'' = \sqrt{\gamma'}\,\mathcal{C}_0\,\mathcal{C}_1$ and $\mathcal{C}_2'' = \sqrt{2/\gamma'}\,\mathcal{C}_2$. Hence, we obtain a similar prediction error bound as the one given in Remark 2.2.2, see (2.6). Notice however that in the model and references given in Remark 2.2.2, the Riemannian diameter of the parameter set Θ_T is bounded by a constant free of T, whereas in this section it grows (sublinearly) with T without degrading the prediction error bound.

2.9 Proofs of Theorems 2.2.1 and 2.2.5

2.9.1 Proof of Theorem 2.2.1

Let us bound the prediction error $\hat{R}_T := \left\|\hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star\Phi_T(\vartheta^\star)\right\|_T$. By definition (2.3) of $\hat{\beta}$ and $\hat{\vartheta}$ for the tuning parameter κ, we have:

$$\frac{1}{2}\left\|y - \hat{\beta}\Phi_T(\hat{\vartheta})\right\|_T^2 + \kappa\|\hat{\beta}\|_{\ell_1} \leq \frac{1}{2}\|y - \beta^\star\Phi_T(\vartheta^\star)\|_T^2 + \kappa\|\beta^\star\|_{\ell_1}.$$

We define the application $\hat{\Upsilon}$ from H_T to $\mathbb{R}$ by:

$$\hat{\Upsilon}(f) = \left\langle \hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star\Phi_T(\vartheta^\star), f\right\rangle_T.$$

This gives, by rearranging terms and using the equation of the model $y = \beta^\star\Phi_T(\vartheta^\star) + w_T$, that:

$$\frac{1}{2}\hat{R}_T^2 \leq \hat{\Upsilon}(w_T) + \kappa\left(\|\beta^\star\|_{\ell_1} - \|\hat{\beta}\|_{\ell_1}\right). \tag{2.53}$$

Next, we shall expand the two terms on the right hand side of (2.53) according to $\hat{\beta}_\ell$ close to some $\beta_k^\star$ or not. In the rest of the proof, we fix $r > 0$ so that Assumptions 2.6.1 and 2.6.2, are verified by $\mathcal{Q}^\star$. In particular, for all $k \neq k'$ in $S^\star = \{k'' \in \{1, \cdots, K\},\, \beta_{k''}^\star \neq 0\}$ we have $\mathfrak{d}_T(\theta_k^\star, \theta_{k'}^\star) > 2r$.

Recall the definitions given in Section 2.2 of the sets of indices $\hat{S}$, $\tilde{S}_k(r)$ and $\tilde{S}(r)$ for $k \in S^\star$. Since the closed balls $\mathcal{B}_T(\theta_k^\star, r)$ with $k \in S^\star$ are pairwise disjoint, the sets $\tilde{S}_k(r)$, for $k \in S^\star$, are also pairwise disjoint and one can write the following decomposition:

$$\begin{aligned}\hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star\Phi_T(\vartheta^\star) &= \sum_{k=1}^{K} \hat{\beta}_k \phi_T(\hat{\theta}_k) - \sum_{k\in S^\star} \beta_k^\star \phi_T(\theta_k^\star) \\ &= \sum_{k\in S^\star} \sum_{\ell\in\tilde{S}_k(r)} \hat{\beta}_\ell \phi_T(\hat{\theta}_\ell) + \sum_{k\in\tilde{S}(r)^c} \hat{\beta}_k \phi_T(\hat{\theta}_k) - \sum_{k\in S^\star} \beta_k^\star \phi_T(\theta_k^\star).\end{aligned}$$

This decomposition groups the elements of the predicted mixture according to the proximity of the estimated parameter $\hat{\theta}_\ell$ to a true underlying parameter $\theta_k^\star$ to be estimated. We use a Taylor-type expansion with the Riemannian metric $\mathfrak{d}_T$ for the function $\phi_T(\theta)$ around the elements of $\mathcal{Q}^\star$. By Assumption 2.3.1, the function ϕ_T is twice continuously differentiable with respect to the variable θ and the function g_T defined in (2.12) is positive on Θ_T and of class $\mathcal{C}^1$ by Assumption 2.3.2. We set in this section $\tilde{D}_{i;T}[\phi_T] = \phi_T^{[i]}$ for $i = 0, 1, 2$. According to Lemma 2.4.2, we have for any $\theta_k^\star$ and $\hat{\theta}_\ell$ in Θ_T:

$$\phi_T(\hat{\theta}_\ell) = \phi_T(\theta_k^\star) + \operatorname{sign}(\hat{\theta}_\ell - \theta_k^\star)\, \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)\, \phi_T^{[1]}(\theta_k^\star) + \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)^2 \int_0^1 (1-s)\phi_T^{[2]}(\gamma_s^{(k\ell)})\, \mathrm{d}s,$$

where $\gamma^{(k\ell)}$ is a distance realizing geodesic path belonging to Θ_T such that $\gamma_0^{(k\ell)} = \theta_k^\star$, $\gamma_1^{(k\ell)} = \hat{\theta}_\ell$ and $\mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star) = \mathcal{L}_T(\gamma^{(k\ell)})$. Hence we obtain:

$$\begin{aligned}\hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star\Phi_T(\vartheta^\star) = \sum_{k\in S^\star} I_{0,k}(r)\, \phi_T(\theta_k^\star) + \sum_{k\in S^\star} I_{1,k}(r)\, \phi_T^{[1]}(\theta_k^\star) + \sum_{k\in\tilde{S}(r)^c} \hat{\beta}_k\, \phi_T(\hat{\theta}_k) \\ + \sum_{k\in S^\star} \left(\sum_{\ell\in\tilde{S}_k(r)} \hat{\beta}_\ell\, \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)^2 \int_0^1 (1-s)\phi_T^{[2]}(\gamma_s^{(k\ell)})\, \mathrm{d}s \right), \quad (2.54)\end{aligned}$$

with

$$I_{0,k}(r) = \Big(\sum_{\ell\in\tilde{S}_k(r)} \hat{\beta}_\ell \Big) - \beta_k^\star \quad \text{and} \quad I_{1,k}(r) = \sum_{\ell\in\tilde{S}_k(r)} \hat{\beta}_\ell\, \operatorname{sign}(\hat{\theta}_\ell - \theta_k^\star)\, \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star).$$

Let us introduce some notations in order to bound the different terms of the expansion above:

$$\begin{aligned} I_0(r) &= \sum_{k\in S^\star} |I_{0,k}(r)| \quad \text{and} \quad I_1(r) = \sum_{k\in S^\star} |I_{1,k}(r)|, \\ I_{2,k}(r) &= \sum_{\ell\in\tilde{S}_k(r)} \left|\hat{\beta}_\ell\right| \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)^2 \quad \text{and} \quad I_2(r) = \sum_{k\in S^\star} I_{2,k}(r), \quad (2.55) \\ I_3(r) &= \sum_{\ell\in\tilde{S}(r)^c} \left|\hat{\beta}_\ell\right| = \left\|\hat{\beta}_{\tilde{S}(r)^c}\right\|_{\ell_1}, \quad (2.56)\end{aligned}$$

and we omit the dependence in r when there is no ambiguity.

We bound the difference $\|\beta^\star\|_{\ell_1} - \|\hat{\beta}\|_{\ell_1}$ by noticing that:

$$\|\beta^\star\|_{\ell_1} - \|\hat{\beta}\|_{\ell_1} = \sum_{k\in S^\star} \Big(|\beta_k^\star| - \sum_{\ell\in\tilde{S}_k(r)} |\hat{\beta}_\ell| \Big) - \sum_{k\in\tilde{S}(r)^c} \left|\hat{\beta}_k\right| \le \sum_{k\in S^\star} \left|\beta_k^\star - \sum_{\ell\in\tilde{S}_k(r)} \hat{\beta}_\ell\right| = I_0. \quad (2.57)$$

In the next lemma, we give an upper bound of I_0. Recall the constants C_N' and C_F from Assumption 2.6.1.

Lemma 2.9.1. *Under the assumptions of Theorem 2.2.1 and with the element $p_1 \in H_T$ from Assumption 2.6.1 associated to the function $v : \mathcal{Q}^\star \to \{-1, 1\}$ defined by:*

$$v(\theta_k^\star) = \operatorname{sign}(I_{0,k}) \quad \textit{for all } k \in S^\star,$$

we get that:

$$I_0 \leq C'_N I_2 + (1 - C_F) I_3 + |\hat{\Upsilon}(p_1)|. \tag{2.58}$$

Proof. Let $v \in \{-1, 1\}^s$ with entries $v_k = v(\theta_k^\star)$ so that:

$$I_0 = \sum_{k \in S^\star} |I_{0,k}| = \sum_{k \in S^\star} v_k I_{0,k} = \sum_{k \in S^\star} v_k \Big(\big(\sum_{\ell \in \tilde{S}_k(r)} \hat{\beta}_\ell \big) - \beta_k^\star \Big).$$

Let p_1 be an element of H_T from Assumption 2.6.1 associated to the application v such that properties *(i)*-*(iv)* therein hold. By adding and substracting $\sum_{k \in S^\star} \sum_{\ell \in \tilde{S}_k(r)} \hat{\beta}_\ell \left\langle \phi_T(\hat{\theta}_\ell), p_1 \right\rangle_T$ to I_0 and using the property *(ii)* satisfied by the element p_1, that is, $\langle \phi_T(\theta_k^\star), p_1 \rangle_T = v_k$ for all $k \in S^\star$, we obtain:

$$I_0 = \sum_{k \in S^\star} \sum_{\ell \in \tilde{S}_k(r)} \hat{\beta}_\ell \left(v_k - \left\langle \phi_T(\hat{\theta}_\ell), p_1 \right\rangle_T \right) + \left\langle \hat{\beta} \Phi_T(\hat{\vartheta}) - \beta^\star \Phi_T(\vartheta^\star), p_1 \right\rangle_T - \sum_{\ell \in \tilde{S}(r)^c} \hat{\beta}_\ell \left\langle \phi_T(\hat{\theta}_\ell), p_1 \right\rangle_T.$$

We deduce that:

$$I_0 \leq \sum_{k \in S^\star} \sum_{\ell \in \tilde{S}_k(r)} |\hat{\beta}_\ell| \left| v_k - \left\langle \phi_T(\hat{\theta}_\ell), p_1 \right\rangle_T \right| + |\hat{\Upsilon}(p_1)| + \sum_{\ell \in \tilde{S}(r)^c} |\hat{\beta}_\ell| \left| \left\langle \phi_T(\hat{\theta}_\ell), p_1 \right\rangle_T \right|.$$

Notice that for $\ell \in \tilde{S}(r)^c$, $\hat{\theta}_\ell \notin \bigcup_{k \in S^\star} \mathcal{B}_T(\theta_k^\star, r)$. Then, by using the properties *(ii)* and *(iii)* from Assumption 2.6.1, we get that (2.58) holds with the constants C'_N and C_F from Assumption 2.6.1. □

In the next lemma, we give an upper bound of I_1. Recall the constants c_N and c_F from Assumption 2.6.2.

Lemma 2.9.2. *Under the assumptions of Theorem 2.2.1 and with the element $q_0 \in H_T$ from Assumption 2.6.2 associated to the function $v : \mathcal{Q}^\star \to \{-1, 1\}$ defined by:*

$$v(\theta_k^\star) = \operatorname{sign}(I_{1,k}) \quad \textit{for all } k \in S^\star,$$

we get that:

$$I_1 \leq c_N I_2 + c_F I_3 + |\hat{\Upsilon}(q_0)|. \tag{2.59}$$

Proof. Let $v \in \{-1, 1\}^s$ with entries $v_k = v(\theta_k^\star)$ so that:

$$I_1 = \sum_{k \in S^\star} |I_{1,k}| = \sum_{k \in S^\star} v_k I_{1,k} = \sum_{k \in S^\star} \sum_{\ell \in \tilde{S}_k(r)} \hat{\beta}_\ell \, v_k \, \operatorname{sign}(\hat{\theta}_\ell - \theta_k^\star) \, \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star).$$

Let $q_0 \in H_T$ from Assumption 2.6.2 associated to the application v such that properties *(i)*-*(iii)* therein hold. By adding and substracting $\sum_{\ell \in \tilde{S}(r)} \hat{\beta}_\ell \left\langle \phi_T(\hat{\theta}_\ell), q_0 \right\rangle_T = \left\langle \hat{\beta} \Phi_T(\hat{\vartheta}), q_0 \right\rangle_T - \sum_{\ell \in \tilde{S}(r)^c} \hat{\beta}_\ell \left\langle \phi_T(\hat{\theta}_\ell), q_0 \right\rangle_T$ to I_1 and using the triangle inequality, we obtain:

$$I_1 \leq \sum_{k \in S^\star} \sum_{\ell \in \tilde{S}_k(r)} |\hat{\beta}_\ell| \left| v_k \, \operatorname{sign}(\hat{\theta}_\ell - \theta_k^\star) \, \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star) - \left\langle \phi_T(\hat{\theta}_\ell), q_0 \right\rangle_T \right| + \sum_{\ell \in \tilde{S}(r)^c} |\hat{\beta}_\ell| \left| \left\langle \phi_T(\hat{\theta}_\ell), q_0 \right\rangle_T \right| + \left| \left\langle \hat{\beta} \Phi_T(\hat{\vartheta}), q_0 \right\rangle_T \right|.$$

The property (*i*) of Assumption 2.6.2 gives that $\langle \phi_T(\theta_k^\star), q_0 \rangle_T = 0$ for all $k \in S^\star$. This implies that $\langle \beta^\star \Phi_T(\vartheta^\star), q_0 \rangle_T = 0$. Then, by using the definition of I_2 and I_3 from (2.55)-(2.56) and the properties (*i*) and (*ii*) of Assumption 2.6.2, we obtain:

$$I_1 \leq c_N I_2 + c_F I_3 + \left| \left\langle \hat{\beta} \Phi_T(\hat{\vartheta}), q_0 \right\rangle_T \right| = c_N I_2 + c_F I_3 + |\hat{\Upsilon}(q_0)|,$$

with the constants c_N and c_F from Assumption 2.6.2. □

We consider the following suprema of Gaussian processes for $i = 0, 1, 2$:

$$M_i = \sup_{\theta \in \Theta_T} \left| \left\langle w_T, \phi_T^{[i]}(\theta) \right\rangle_T \right|.$$

By using the expansion (2.54) and the bounds (2.59) and (2.58) for the second inequality, we obtain:

$$\begin{aligned} |\hat{\Upsilon}(w_T)| &\leq (I_0 + I_3) M_0 + I_1 M_1 + I_2 \, 2^{-1} \, M_2 && (2.60) \\ &\leq (C'_N I_2 + (2 - C_F) I_3 + |\hat{\Upsilon}(p_1)|) M_0 \\ &\quad + (c_N I_2 + c_F I_3 + |\hat{\Upsilon}(q_0)|) M_1 + I_2 \, 2^{-1} \, M_2. && (2.61) \end{aligned}$$

At this point, one needs to bound I_2 and I_3. In order to do so, we will bound from above and from below the Bregman divergence D_B defined by:

$$D_B = \|\hat{\beta}\|_{\ell_1} - \|\beta^\star\|_{\ell_1} - \hat{\Upsilon}(p_0), \tag{2.62}$$

where p_0 is the element of H_T given by the Assumption 2.6.1 associated to the application $v : \mathcal{Q}^\star \to \{-1, 1\}$ given by:

$$v(\theta_k^\star) = \operatorname{sign}(\beta_k^\star) \quad \text{for all } k \in S^\star. \tag{2.63}$$

The next lemma gives a lower bound of the Bregman divergence.

Lemma 2.9.3. *Under the assumptions of Theorem 2.2.1 and with the constants C_N and C_F of Assumption 2.6.1, we get that:*

$$D_B \geq C_N I_2 + C_F I_3. \tag{2.64}$$

Proof. By definition (2.62) of D_B we have:

$$D_B = \sum_{k \in \hat{S}} |\hat{\beta}_k| - \hat{\beta}_k \left\langle \phi_T(\hat{\theta}_k), p_0 \right\rangle_T - \left(\sum_{k \in S^\star} |\beta_k^\star| - \beta_k^\star \langle \phi_T(\theta_k^\star), p_0 \rangle_T \right).$$

By using the interpolating properties of the element p_0 of H_T from Assumption 2.6.1 associated to the function v defined in (2.63), we have $\sum_{k \in S^\star} |\beta_k^\star| - \beta_k^\star \langle \phi_T(\theta_k^\star), p_0 \rangle_T = 0$. Hence, we deduce that:

$$\begin{aligned} D_B &= \sum_{k \in \hat{S}} |\hat{\beta}_k| - \hat{\beta}_k \left\langle \phi_T(\hat{\theta}_k), p_0 \right\rangle_T \\ &\geq \sum_{k \in \hat{S}} |\hat{\beta}_k| - |\hat{\beta}_k| \left| \left\langle \phi_T(\hat{\theta}_k), p_0 \right\rangle_T \right| \\ &= \sum_{\ell \in \tilde{S}(r)} |\hat{\beta}_\ell| \left(1 - \left| \left\langle \phi_T(\hat{\theta}_\ell), p_0 \right\rangle_T \right| \right) + \sum_{k \in \tilde{S}(r)^c} |\hat{\beta}_k| \left(1 - \left| \left\langle \phi_T(\hat{\theta}_k), p_0 \right\rangle_T \right| \right). \end{aligned}$$

Thanks to properties (*i*) and (*iii*) of Assumption 2.6.1 and the definitions (2.55) and (2.56) of I_2 and I_3, we obtain:

$$D_B \geq \sum_{k \in S^\star} \sum_{\ell \in \tilde{S}_k(r)} C_N |\hat{\beta}_\ell| \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)^2 + \sum_{k \in \tilde{S}(r)^c} C_F |\hat{\beta}_k| = C_N I_2 + C_F I_3,$$

where the constants C_N and C_F are that of Assumption 2.6.1. □

We now give an upper bound of the Bregman divergence.

Lemma 2.9.4. *Under the assumptions of Theorem 2.2.1, we have:*

$$\kappa D_B \leq I_2\left(C'_N M_0 + c_N M_1 + 2^{-1} M_2\right) + I_3\left((2 - C_F)M_0 + c_F M_1\right) + |\hat{\Upsilon}(p_1)|M_0 + |\hat{\Upsilon}(q_0)|M_1 + \kappa|\hat{\Upsilon}(p_0)|. \quad (2.65)$$

Proof. Recall that $\mathcal{Q}^\star \subset \Theta_T$. We deduce from (2.53) that:

$$\kappa(||\hat{\beta}||_{\ell_1} - ||\beta^\star||_{\ell_1}) \leq \hat{\Upsilon}(w_T) - \frac{1}{2}\left\|\beta^\star \Phi_T(\vartheta^\star) - \hat{\beta}\Phi_T(\hat{\vartheta})\right\|_T^2 \leq \hat{\Upsilon}(w_T).$$

Using (2.62), we obtain:

$$\kappa D_B \leq |\hat{\Upsilon}(w_T)| + \kappa|\hat{\Upsilon}(p_0)|.$$

Then, use (2.61) to get (2.65). □

By combining the upper and lower bounds (2.64) and (2.65), we deduce that:

$$I_2\left(C_N - \frac{1}{\kappa}\left(C'_N M_0 + c_N M_1 + 2^{-1} M_2\right)\right) + I_3\left(C_F - \frac{1}{\kappa}\left((2 - C_F)M_0 + c_F M_1\right)\right) \leq \frac{1}{\kappa}|\hat{\Upsilon}(p_1)|M_0 + \frac{1}{\kappa}|\hat{\Upsilon}(q_0)|M_1 + |\hat{\Upsilon}(p_0)|. \quad (2.66)$$

We define the events:

$$\mathcal{A}_i = \{M_i \leq \mathcal{C}\,\kappa\}, \quad \text{for } i \in \{0,1,2\} \quad \text{and} \quad \mathcal{A} = \mathcal{A}_0 \cap \mathcal{A}_1 \cap \mathcal{A}_2, \quad (2.67)$$

where:

$$\mathcal{C} = \frac{C_F}{2(2 - C_F + c_F)} \wedge \frac{C_N}{2(C'_N + c_N + 2^{-1})}.$$

(We shall prove in (2.76) that the event $\mathcal{A}$ occurs with high probability.) We get from Inequality (2.66), that on the event $\mathcal{A}$:

$$C_N I_2 + C_F I_3 \leq 2\mathcal{C}'\left(|\hat{\Upsilon}(p_1)| + |\hat{\Upsilon}(q_0)| + |\hat{\Upsilon}(p_0)|\right) \quad \text{with} \quad \mathcal{C}' = \mathcal{C} \vee 1. \quad (2.68)$$

By reinjecting (2.57), (2.61), (2.58) and (2.59) in (2.53) one gets:

$$\frac{1}{2}\hat{R}_T^2 \leq I_2(C'_N M_0 + c_N M_1 + 2^{-1} M_2 + \kappa C'_N) + I_3((2 - C_F)M_0 + c_F M_1 + \kappa(1 - C_F)) + |\hat{\Upsilon}(p_1)|(M_0 + \kappa) + |\hat{\Upsilon}(q_0)|M_1.$$

Using (2.68), we obtain an upper bound for the prediction error on the event $\mathcal{A}$:

$$\hat{R}_T^2 \leq C\,\kappa\,(|\hat{\Upsilon}(p_0)| + |\hat{\Upsilon}(p_1)| + |\hat{\Upsilon}(q_0)|), \quad (2.69)$$

with

$$C = 4\mathcal{C}'\left(1 + \frac{\mathcal{C}'}{C_N}(2C'_N + c_N + 1) + \frac{\mathcal{C}'}{C_F}(3 - 2C_F + c_F)\right).$$

Using the Cauchy-Schwarz inequality and the definition of $\hat{\Upsilon}$, we get that for $f \in H_T$:

$$|\hat{\Upsilon}(f)| \leq \hat{R}_T \|f\|_T. \quad (2.70)$$

Using Assumption 2.6.1 *(iv)* for p_0 and p_1, and Assumption 2.6.2 *(iii)* for q_0, we get:

$$\|p_0\|_T \leq C_B\sqrt{s}, \quad \|p_1\|_T \leq C_B\sqrt{s} \quad \text{and} \quad \|q_0\|_T \leq c_B\sqrt{s}. \tag{2.71}$$

Plugging this in (2.69), we get that on the event $\mathcal{A}$:

$$\hat{R}_T^2 \leq \mathcal{C}_0\,\kappa\hat{R}_T\,\sqrt{s} \quad \text{with} \quad \mathcal{C}_0 = (c_B + 2C_B)C. \tag{2.72}$$

This gives (2.4).

The proof of (2.5) is postponed to Section 2.9.2 and will be easily deduced from the first and third inequalities in (2.7).

To complete the proof of Theorem 2.2.1 we shall give a lower bound for the probability of the event $\mathcal{A}$ defined in (2.67). For $i = 0, 1, 2$ and $\theta \in \Theta$, set $X_i(\theta) = \left\langle w_T, \phi_T^{[i]}(\theta)\right\rangle_T$ a real centered Gaussian process with continuously differentiable sample paths, so that its supremum is $M_i = \sup_{\Theta_T} |X_i|$.

We first consider $i = 0$. We have, thanks to (2.28) and (2.25) for the second part:

$$\|\phi_T(\theta)\|_T^2 = 1 \quad \text{and} \quad \left\|\phi_T^{[1]}(\theta)\right\|_T^2 = \mathcal{K}_T^{[1,1]}(\theta,\theta) = 1.$$

Recall Assumption 2.1.1 on the noise w_T holds. We deduce from Lemma 2.11.1 with $C_1 = C_2 = 1$ that:

$$\mathbb{P}\left(\mathcal{A}_0^c\right) = \mathbb{P}\left(\sup_{\Theta_T}|X_0| > \mathcal{C}\,\kappa\right) \leq c_0\left(\sigma\frac{|\Theta_T|_{\mathfrak{d}_T}\sqrt{\Delta_T}}{\mathcal{C}\,\kappa} \vee 1\right)\mathrm{e}^{-(\mathcal{C}\,\kappa)^2/(4\sigma^2\Delta_T)}, \tag{2.73}$$

where $|\Theta_T|_{\mathfrak{d}_T}$ denotes the diameter of the set Θ_T with respect to the metric $\mathfrak{d}_T$ and $c_0 = 3$.

We consider $i = 1$. Thanks to (2.28), we get:

$$\left\|\phi_T^{[1]}(\theta)\right\|_T^2 = 1 \quad \text{and} \quad \left\|\tilde{D}_{1;T}[\phi_T^{[1]}](\theta)\right\|_T^2 = \left\|\phi_T^{[2]}(\theta)\right\|_T^2 = \mathcal{K}_T^{[2,2]}(\theta,\theta).$$

Recall $L_{2,2}$ and $\mathcal{V}_T$ are defined in (2.31) and (2.34). Since Assumptions 2.5.1 and 2.5.2 hold, we get that for $\theta \in \Theta_T$:

$$\mathcal{K}_T^{[2,2]}(\theta,\theta) \leq L_{2,2} + \mathcal{V}_T \leq 2L_{2,2}.$$

We deduce from Lemma 2.11.1 with $C_1 = 1$ and $C_2 = \sqrt{2L_{2,2}}$ and taking $c_1 = 2\sqrt{2L_{2,2}} + 1$, that:

$$\mathbb{P}\left(\mathcal{A}_1^c\right) = \mathbb{P}\left(\sup_{\Theta_T}|X_1| > \mathcal{C}\,\kappa\right) \leq c_1\left(\sigma\frac{|\Theta_T|_{\mathfrak{d}_T}\sqrt{\Delta_T}}{\mathcal{C}\,\kappa} \vee 1\right)\mathrm{e}^{-(\mathcal{C}\,\kappa)^2/(4\sigma^2\Delta_T)}. \tag{2.74}$$

We consider $i = 2$. Thanks to (2.28), we get:

$$\left\|\phi_T^{[2]}(\theta)\right\|_T^2 = \mathcal{K}_T^{[2,2]}(\theta,\theta) \quad \text{and} \quad \left\|\tilde{D}_{1;T}[\phi_T^{[2]}](\theta)\right\|_T^2 = \left\|\phi_T^{[3]}(\theta)\right\|_T^2 = \mathcal{K}_T^{[3,3]}(\theta,\theta).$$

Recall the definition of the function h_∞ given in (2.30) and the constants $L_{2,2}$, L_3, $\mathcal{V}_T$ defined in (2.31) and (2.34). Using also Assumption 2.5.2 so that $\mathcal{V}_T \leq L_{2,2} \wedge L_3$, we get that for all $\theta \in \Theta_T$:

$$\mathcal{K}_T^{[2,2]}(\theta,\theta) \leq L_{2,2} + \mathcal{V}_T \leq 2L_{2,2} \quad \text{and} \quad \mathcal{K}_T^{[3,3]}(\theta,\theta) \leq L_3 + \mathcal{V}_T \leq 2\,L_3.$$

We deduce from Lemma 2.11.1 with $C_1 = \sqrt{2L_{2,2}}$ and $C_2 = \sqrt{2L_3}$ and taking $c_2 = 2\sqrt{2L_3}+1$, that:

$$\mathbb{P}\left(\mathcal{A}_2^c\right) = \mathbb{P}\left(\sup_{\Theta_T}|X_2| > \mathcal{C}\,\kappa\right) \leq c_2\left(\sigma\frac{|\Theta_T|_{\mathfrak{d}_T}\sqrt{\Delta_T}}{\mathcal{C}\,\kappa} \vee 1\right)\mathrm{e}^{-(\mathcal{C}\,\kappa)^2/(8\sigma^2\Delta_T L_{2,2})}. \tag{2.75}$$

Since $\mathcal{A} = \mathcal{A}_0 \cap \mathcal{A}_1 \cap \mathcal{A}_2$, we deduce from (2.73), (2.74) and (2.75) that:

$$\mathbb{P}\left(\mathcal{A}^c\right) = \mathbb{P}\left(\mathcal{A}_0^c \cup \mathcal{A}_1^c \cup \mathcal{A}_2^c\right) \leq \mathcal{C}_2' \left(\sigma \frac{|\Theta_T|_{\mathfrak{d}_T} \sqrt{\Delta_T}}{\mathcal{C}\kappa} \vee 1\right) \mathrm{e}^{-\kappa^2/(\mathcal{C}_1^2 \sigma^2 \Delta_T)},$$

with the finite positive constants:

$$\mathcal{C}_1 = \frac{2}{\mathcal{C}} \left(1 \vee \sqrt{2L_{2,2}}\right) \quad \text{and} \quad \mathcal{C}_2' = c_0 + c_1 + c_2.$$

By taking $\kappa \geq \mathcal{C}_1 \sigma \sqrt{\Delta_T \log \tau}$, for any positive constant $\tau > 1$, we get:

$$\mathbb{P}\left(\mathcal{A}_0^c \cup \mathcal{A}_1^c \cup \mathcal{A}_2^c\right) \leq \mathcal{C}_2 \left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau\sqrt{\log \tau}} \vee \frac{1}{\tau}\right) \quad \text{with} \quad \mathcal{C}_2 = \mathcal{C}_2' \left(\frac{1}{\mathcal{C}\mathcal{C}_1} \vee 1\right). \tag{2.76}$$

This completes the proof of the theorem.

2.9.2 Proof of Theorem 2.2.5 and of Equation (2.5)

We keep notations from Section 2.9.1. Recall that Assumptions (i)-(v) of Theorem 2.2.1 are in force. We shall first provide an upper bound of I_i for $i = 0, 1, 2, 3$. We deduce from (2.70), (2.71) and (2.72), that, on the event $\mathcal{A}$:

$$|\hat{\Upsilon}(p_0)| \leq \mathcal{C}_0 C_B \kappa s, \quad |\hat{\Upsilon}(p_1)| \leq \mathcal{C}_0 C_B \kappa s \quad \text{and} \quad |\hat{\Upsilon}(q_0)| \leq \mathcal{C}_0 c_B \kappa s.$$

Then, we obtain from (2.68) that, on the event $\mathcal{A}$:

$$I_3 \leq \mathcal{C}_5 \kappa s \quad \text{and} \quad I_2 \leq \mathcal{C}_6 \kappa s \quad \text{with} \quad \mathcal{C}_5 = 2\frac{\mathcal{C}'}{C_F}\mathcal{C}_0(c_B + 2C_B) \quad \text{and} \quad \mathcal{C}_6 = \frac{C_F}{C_N}\mathcal{C}_5. \tag{2.77}$$

This gives the third inequality in (2.7), as well as Inequality (2.8) in Remark 2.2.6. We also deduce from (2.58) that, on the event $\mathcal{A}$:

$$I_0 \leq \mathcal{C}_4 \kappa s \quad \text{with} \quad \mathcal{C}_4 = C_N' \mathcal{C}_6 + (1 - C_F)\mathcal{C}_5 + \mathcal{C}_0 C_B. \tag{2.78}$$

This gives the second inequality in (2.7).

We now establish the first inequality in (2.7). We deduce from (2.53) that:

$$\kappa(\|\hat{\beta}\|_{\ell_1} - \|\beta^\star\|_{\ell_1}) \leq \hat{\Upsilon}(w_T). \tag{2.79}$$

Then, using the bounds (2.78) and (2.77) on I_0, I_2 and I_3, we deduce from (2.60) and (2.59) that, on the event $\mathcal{A}$:

$$|\hat{\Upsilon}(w_T)| \leq \mathcal{C}_7 s \kappa^2 \quad \text{with} \quad \mathcal{C}_7 = \mathcal{C}\left(\mathcal{C}_4 + \mathcal{C}_5(1 + c_F) + \mathcal{C}_6(1 + c_N) + \mathcal{C}_0 c_B\right). \tag{2.80}$$

Thus, (2.79) and (2.80) imply that, on the event $\mathcal{A}$:

$$\|\hat{\beta}\|_{\ell_1} - \|\beta^\star\|_{\ell_1} \leq \mathcal{C}_7 s \kappa. \tag{2.81}$$

Then, use (2.57) and (2.78) to deduce that, on the event $\mathcal{A}$:

$$\left|\|\hat{\beta}\|_{\ell_1} - \|\beta^\star\|_{\ell_1}\right| \leq (\mathcal{C}_4 \vee \mathcal{C}_7) s \kappa.$$

This proves (2.5) (we shall take $\mathcal{C}_3 = \mathcal{C}_7 + 2\mathcal{C}_4$, see below). Let $\mathcal{I}^+$ (resp. $\mathcal{I}^-$) be the set of indices $k \in S^\star$ such that the quantity $\left(\sum_{\ell \in \tilde{S}_k(r)} |\hat{\beta}_\ell|\right) - |\beta_k^\star|$ is non negative (resp. negative). We have the following decomposition:

$$\begin{aligned}
\sum_{k \in S^\star} \Big| \sum_{\ell \in \tilde{S}_k(r)} |\hat{\beta}_\ell| - |\beta_k^\star| \Big| &= \sum_{k \in \mathcal{I}^+} \Big(\sum_{\ell \in \tilde{S}_k(r)} |\hat{\beta}_\ell| - |\beta_k^\star| \Big) + \sum_{k \in \mathcal{I}^-} \Big(|\beta_k^\star| - \sum_{\ell \in \tilde{S}_k(r)} |\hat{\beta}_\ell| \Big) \\
&\leq \|\hat{\beta}\|_{\ell_1} - \|\beta^\star\|_{\ell_1} + 2 \sum_{k \in \mathcal{I}^-} \Big(|\beta_k^\star| - \sum_{\ell \in \tilde{S}_k(r)} |\hat{\beta}_\ell| \Big) \\
&\leq \|\hat{\beta}\|_{\ell_1} - \|\beta^\star\|_{\ell_1} + 2I_0.
\end{aligned}$$

Then, use (2.78) and (2.81) to obtain the first inequality (2.7) with $\mathcal{C}_3 = \mathcal{C}_7 + 2\mathcal{C}_4$. This ends the proof of Theorem 2.2.5.

2.10 Construction of certificate functions

2.10.1 Proof of Proposition 2.7.4 (Construction of an interpolating certificate)

This section is devoted to the proof of Proposition 2.7.4. We closely follow the proof of [Poon et al., 2021] taking into account the approximation of the kernel $\mathcal{K}_T$ by the kernel $\mathcal{K}_\infty$, which is measured through the quantity $\mathcal{V}_T$ defined in (2.34).

Let $T \in \mathbb{N}$ and $s \in \mathbb{N}^*$. Recall Assumptions 2.3.2 (and thus 2.3.1 on the regularity of φ_T) and 2.5.1 on the regularity of the asymptotic kernel $\mathcal{K}_\infty$ are in force. Let $\rho \geq 1$, let $r \in (0, 1/\sqrt{2L_{0,2}})$ and $u_\infty \in \left(0, H_\infty^{(2)}(r,\rho)\right)$ such that *(ii)*, *(iii)*, *(iv)* and *(v)* of Proposition 2.7.4 hold. We denote by $\|\cdot\|_{\mathrm{op}}$ the operator norm associated to the ℓ_∞ norm on $\mathbb{R}^s$.

By assumption $\delta_\infty(u_\infty, s)$ is finite. Let $\vartheta^\star = (\theta_1^\star, \dots, \theta_s^\star) \in \Theta^s_{T, 2\rho_T\, \delta_\infty(u_\infty,s)}$. We note $\mathcal{Q}^\star = \{\theta_i^\star,\, 1 \leq i \leq s\}$ the set of parameters of cardinal s. By Lemma 2.7.3, we have:

$$\Theta^s_{T,\rho_T\delta_\infty(u_\infty,s)} \subseteq \Theta^s_{T,\delta_T(u_T(s),s)} \quad \text{where} \quad u_T(s) = u_\infty + (s-1)\mathcal{V}_T.$$

Hence we have:

$$\vartheta^\star \in \Theta^s_{T,\delta_T(u_T(s),s)}. \tag{2.82}$$

Set

$$\Gamma^{[i,j]} = \mathcal{K}_T^{[i,j]}(\vartheta^\star) \quad \text{and} \quad \Gamma = \begin{pmatrix} \Gamma^{[0,0]} & \Gamma^{[1,0]\top} \\ \Gamma^{[1,0]} & \Gamma^{[1,1]} \end{pmatrix}. \tag{2.83}$$

We deduce from (2.39) and (2.82) that:

$$\left\|I - \Gamma^{[0,0]}\right\|_{\mathrm{op}} \leq u_T(s),\ \left\|I - \Gamma^{[1,1]}\right\|_{\mathrm{op}} \leq u_T(s),\ \left\|\Gamma^{[1,0]}\right\|_{\mathrm{op}} \leq u_T(s) \text{ and } \left\|\Gamma^{[1,0]\top}\right\|_{\mathrm{op}} \leq u_T(s). \tag{2.84}$$

For simplicity, for an expression A we write A_T for $A_{\mathcal{K}_T}$. Using this convention, recall the definition of the derivative operator $\tilde{D}_{i;T}$ and write $\phi_T^{[1]}$ for $\tilde{D}_{1;T}[\phi_T]$.

Let $\alpha = (\alpha_1, \dots, \alpha_s)^\top$ and $\xi = (\xi_1, \dots, \xi_s)^\top$ be elements of $\mathbb{R}^s$. Let $p_{\alpha,\xi}$ be an element of H_T defined by:

$$p_{\alpha,\xi} = \sum_{k=1}^s \alpha_k \phi_T(\theta_k^\star) + \sum_{k=1}^s \xi_k\, \phi_T^{[1]}(\theta_k^\star), \tag{2.85}$$

and, using (2.28) in Lemma 2.4.3, set the interpolating real-valued function $\eta_{\alpha,\xi}$ defined on Θ by:

$$\eta_{\alpha,\xi}(\theta) = \langle \phi_T(\theta), p_{\alpha,\xi}\rangle_T = \sum_{k=1}^s \alpha_k\, \mathcal{K}_T(\theta, \theta_k^\star) + \sum_{k=1}^s \xi_k\, \mathcal{K}_T^{[0,1]}(\theta, \theta_k^\star). \tag{2.86}$$

By Assumption 2.3.2 on the regularity of φ_T and the positivity of g_T and Lemma 2.4.3, we get that the function $\eta_{\alpha,\xi}$ is of class $\mathcal{C}^3$ on Θ, and using (2.20), we get that:

$$\eta_{\alpha,\xi}^{[1]} := \tilde{D}_{1;T}[\eta_{\alpha,\xi}](\theta) = \sum_{k=1}^s \alpha_k\, \mathcal{K}_T^{[1,0]}(\theta, \theta_k^\star) + \sum_{k=1}^s \xi_k\, \mathcal{K}_T^{[1,1]}(\theta, \theta_k^\star). \tag{2.87}$$

We give a preliminary technical lemma.

Lemma 2.10.1. *Let $v = (v_1, \cdots, v_s)^\top \in \{-1, 1\}^s$ be a sign vector. Assume that (2.84) holds with $u_T(s) < 1/2$. Under Assumption 2.3.2, there exist unique $\alpha, \xi \in \mathbb{R}^s$ such that:*

$$\eta_{\alpha,\xi}(\theta_k^\star) = v_k \in \{-1, 1\} \quad \textit{and} \quad \eta_{\alpha,\xi}^{[1]}(\theta_k^\star) = 0 \quad \textit{for} \quad 1 \leq k \leq s. \tag{2.88}$$

Furthermore, we have:

$$\|\alpha\|_{\ell_\infty} \leq \frac{1 - u_T(s)}{1 - 2u_T(s)}, \quad \|\alpha - v\|_{\ell_\infty} \leq \frac{u_T(s)}{1 - 2u_T(s)} \quad \textit{and} \quad \|\xi\|_{\ell_\infty} \leq \frac{u_T(s)}{1 - 2u_T(s)}.$$

Proof of Lemma 2.10.1 . Thanks to (2.28), (2.25) and (2.87), we have:

$$\left(\eta_{\alpha,\xi}(\theta_1^\star),\dots,\eta_{\alpha,\xi}(\theta_s^\star),\eta^{[1]}_{\alpha,\xi}(\theta_1^\star),\dots,\eta^{[1]}_{\alpha,\xi}(\theta_s^\star)\right)^\top = \Gamma\begin{pmatrix}\alpha\\ \xi\end{pmatrix}.$$

Thus, solving (2.88) is equivalent to solving,

$$\Gamma\begin{pmatrix}\alpha\\ \xi\end{pmatrix} = \begin{pmatrix}v\\ 0_s\end{pmatrix},$$

with 0_s the vector of size s with all its components equal to zero.

We first show that Γ is non singular so that α and ξ exist and are uniquely defined. Using Lemma 2.11.3 based on the Schur complement, Γ has an inverse provided that $\Gamma^{[1,1]}$ and $\Gamma_{SC} := \Gamma^{[0,0]} - \Gamma^{[1,0]\top}[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}$ are non singular. We recall that if M is a matrix such that, $\|I-M\|_{\text{op}} < 1$, then M is non singular, $M^{-1} = \sum_{i\geq 0}(I-M)^i$ and $\|M^{-1}\|_{\text{op}} \leq \left(1-\|I-M\|_{\text{op}}\right)^{-1}$.

Recall that by assumption $u_T(s) \leq 1/2$. Then, the second inequality in (2.84) imply that $\left\|I-\Gamma^{[1,1]}\right\|_{\text{op}} < 1$ and thus $\Gamma^{[1,1]}$ is non singular. We now prove that Γ_{SC} is also non singular. Using the triangle inequality we have:

$$\begin{aligned}\|I-\Gamma_{SC}\|_{\text{op}} &= \left\|I-\Gamma^{[0,0]}+\Gamma^{[1,0]\top}[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}\right\|_{\text{op}}\\ &\leq \left\|I-\Gamma^{[0,0]}\right\|_{\text{op}} + \left\|\Gamma^{[1,0]\top}[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}\right\|_{\text{op}}.\end{aligned}$$

Let us bound the terms on the right hand side of the inequality above.
To bound $\left\|\Gamma^{[1,0]\top}[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}\right\|_{\text{op}}$ notice that:

$$\left\|\Gamma^{[1,0]\top}[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}\right\|_{\text{op}} \leq ||\Gamma^{[1,0]}||_{\text{op}}\left\|\Gamma^{[1,0]\top}\right\|_{\text{op}}\left\|[\Gamma^{[1,1]}]^{-1}\right\|_{\text{op}}.$$

We have, thanks to (2.84) for the second inequality:

$$\left\|[\Gamma^{[1,1]}]^{-1}\right\|_{\text{op}} \leq \frac{1}{1-\left\|I-\Gamma^{[1,1]}\right\|_{\text{op}}} \leq \frac{1}{1-u_T(s)}. \tag{2.89}$$

Using (2.84), we get:

$$||I-\Gamma_{SC}||_{\text{op}} \leq u_T(s) + \frac{u_T(s)^2}{1-u_T(s)} = \frac{u_T(s)}{1-u_T(s)}.$$

By assumption, we have $u_T(s) \leq H^{(2)}_\infty(r,\rho) < 1/2$. Hence, we have $\frac{u_T(s)}{1-u_T(s)} < 1$ and thus, Γ_{SC} is non singular. Furthermore, we get:

$$||\Gamma_{SC}^{-1}||_{\text{op}} \leq \frac{1}{1-\|I-\Gamma_{SC}\|_{\text{op}}} \leq \frac{1-u_T(s)}{1-2u_T(s)}. \tag{2.90}$$

As the matrices $\Gamma^{[1,1]}$ and Γ_{SC} are non singular, we deduce that the matrix Γ is non singular.

We now give bounds related to α and ξ. The Lemma 2.11.3 on the Schur complement gives also that:

$$\alpha = \Gamma_{SC}^{-1}v \quad \text{and} \quad \xi = -[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}\Gamma_{SC}^{-1}v.$$

Hence, we deduce that:

$$\|\alpha\|_{\ell_\infty} \leq \left\|\Gamma_{SC}^{-1}\right\|_{\mathrm{op}} \|v\|_{\ell_\infty} \leq \frac{1-u_T(s)}{1-2u_T(s)},$$
$$\|\xi\|_{\ell_\infty} \leq \left\|[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}\Gamma_{SC}^{-1}\right\|_{\mathrm{op}} \|v\|_{\ell_\infty} \leq \left\|[\Gamma^{[1,1]}]^{-1}\right\|_{\mathrm{op}} \left\|\Gamma^{[1,0]}\right\|_{\mathrm{op}} \left\|\Gamma_{SC}^{-1}\right\|_{\mathrm{op}} \leq \frac{u_T(s)}{1-2u_T(s)},$$
$$\|\alpha - v\|_{\ell_\infty} \leq \left\|(\Gamma_{SC}^{-1} - I)\right\|_{\mathrm{op}} \|v\|_{\ell_\infty} \leq \|\Gamma_{SC} - I\|_{\mathrm{op}} \left\|\Gamma_{SC}^{-1}\right\|_{\mathrm{op}} \leq \frac{u_T(s)}{1-2u_T(s)}.$$

This finishes the proof. □

We now fix a sign vector $v = (v_1, \cdots, v_s)^\top \in \{-1, 1\}^s$ and consider $p_{\alpha,\xi}$ and $\eta_{\alpha,\xi}$ with α and ξ characterized by (2.88) from Lemma 2.10.1. Let $e_\ell \in \mathbb{R}^s$ be the vector with all the entries equal to zero but the ℓ-th which is equal to 1.

Proof of *(iii)* from Assumption 2.6.1 with $C_F = \varepsilon_\infty(r/\rho)/10$. Let $\theta \in \Theta_T$ such that $\mathfrak{d}_T(\theta, \mathcal{Q}^\star) > r$ (far region). It is enough to prove that $|\eta_{\alpha,\xi}(\theta)| \leq 1 - C_F$. Let $\theta_\ell^\star$ be one of the elements of $\mathcal{Q}^\star$ closest to θ in terms of the metric $\mathfrak{d}_T$. Since $\vartheta^\star \in \Theta^s_{T,2\rho_T\delta_\infty(u_\infty,s)}$, we have, by the triangle inequality that for any $k \neq \ell$:

$$2\rho_T\,\delta_\infty(u_\infty, s) < \mathfrak{d}_T(\theta_\ell^\star, \theta_k^\star) \leq \mathfrak{d}_T(\theta_\ell^\star, \theta) + \mathfrak{d}_T(\theta, \theta_k^\star) \leq 2\mathfrak{d}_T(\theta, \theta_k^\star).$$

Hence, we have $\vartheta^\star_{\ell,\theta} \in \Theta^s_{T,\rho_T\delta_\infty(u_\infty,s)}$, where $\vartheta^\star_{\ell,\theta}$ denotes the vector $\vartheta^\star$ whose ℓ-th coordinate has been replaced by θ. Then, we obtain from Lemma 2.7.3 that $\Theta^s_{T,\rho_T\delta_\infty(u_\infty,s)} \subseteq \Theta^s_{T,\delta_T(u_T(s),s)}$ and thus:

$$\vartheta^\star_{\ell,\theta} \in \Theta^s_{T,\delta_T(u_T(s),s)}. \tag{2.91}$$

We denote by $\Gamma_{\ell,\theta}$ (resp. $\Gamma^{[i,j]}_{\ell,\theta}$) the matrix Γ (resp. $\Gamma^{[i,j]}$) in (2.83) where $\vartheta^\star$ has been replaced by $\vartheta^\star_{\ell,\theta}$. Notice the upper bounds (2.84) also hold for $\Gamma_{\ell,\theta}$ because of (2.91). Recall we have Equalities (2.29) on the diagonal of the kernel $\mathcal{K}_T$ and its derivatives. Elementary calculations give with $\eta_{\alpha,\xi}$ from Lemma 2.10.1 that:

$$\eta_{\alpha,\xi}(\theta) = e_\ell^\top \left(\Gamma^{[0,0]}_{\ell,\theta} - I\right)\alpha + \mathcal{K}_T(\theta, \theta_\ell^\star)\alpha_\ell + e_\ell^\top \Gamma^{[1,0]\top}_{\ell,\theta}\xi + \mathcal{K}_T^{[0,1]}(\theta, \theta_\ell^\star)\xi_\ell. \tag{2.92}$$

We deduce that:

$$|\eta_{\alpha,\xi}(\theta)| \leq \left\|\Gamma^{[0,0]}_{\ell,\theta} - I\right\|_{\mathrm{op}} \|\alpha\|_{\ell_\infty} + \|\alpha\|_{\ell_\infty} |\mathcal{K}_T(\theta, \theta_\ell^\star)| + \left\|\Gamma^{[1,0]\top}_{\ell,\theta}\right\|_{\mathrm{op}} \|\xi\|_{\ell_\infty} + |\mathcal{K}_T^{[0,1]}(\theta, \theta_\ell^\star)| \|\xi\|_{\ell_\infty}.$$

Since θ belongs to the "far region", we have by definition of $\varepsilon_T(r)$ given in (2.35) that:

$$|\mathcal{K}_T(\theta, \theta_\ell^\star)| \leq 1 - \varepsilon_T(r).$$

The triangle inequality, the definitions (2.34) of $\mathcal{V}_T$ and (2.31) of $L_{1,0}$, give:

$$|\mathcal{K}_T^{[0,1]}(\theta, \theta_\ell^\star)| \leq L_{0,1} + \mathcal{V}_T. \tag{2.93}$$

Then, using (2.84) (which holds for $\Gamma_{\ell,\theta}$ thanks to (2.91)), we get that:

$$|\eta_{\alpha,\xi}(\theta)| \leq 1 - \varepsilon_T(r) + \frac{u_T(s)}{1-2u_T(s)}\left(2 + L_{1,0} + \mathcal{V}_T\right).$$

Notice that the function $r \mapsto \varepsilon_\infty(r)$ is increasing. Since $\rho_T \leq \rho$, we get by Lemma 2.7.1 that:

$$\varepsilon_T(r) \geq \varepsilon_\infty(r/\rho_T) - \mathcal{V}_T \geq \varepsilon_\infty(r/\rho) - \mathcal{V}_T.$$

By assumption, we have $u_T(s) \leq H_\infty^{(2)}(r, \rho) \leq 1/4$. Hence, we have $\frac{1}{1-2u_T(s)} \leq 2$. We also have $\mathcal{V}_T \leq 1/2$. Therefore, we get:

$$|\eta_{\alpha,\xi}(\theta)| \leq 1 - \varepsilon_\infty(r/\rho) + \mathcal{V}_T + u_T(s)\left(5 + 2L_{1,0}\right).$$

The assumption $u_T(s) \leq H_\infty^{(2)}(r,\rho)$ gives:

$$u_T(s) \leq \frac{8}{10\,(5+2L_{1,0})}\varepsilon_\infty(r/\rho)\cdot$$

The assumption $\mathcal{V}_T \leq H_\infty^{(1)}(r,\rho)$ gives $\mathcal{V}_T \leq \varepsilon_\infty(r/\rho)/10$. Hence, we have $|\eta_{\alpha,\xi}(\theta)| \leq 1 - \frac{\varepsilon_\infty(r/\rho)}{10}$. Thus, Property *(iii)* from Assumption 2.6.1 holds with $C_F = \varepsilon_\infty(r/\rho)/10$.

Proof of *(i)* from Assumption 2.6.1 with $C_N = \nu_\infty(\rho r)/180$. Let $\theta \in \Theta_T$ such that $\mathfrak{d}_T(\theta, \mathcal{Q}^\star) \leq r$. Let $\ell \in \{1,\cdots,s\}$ such that $\theta \in \mathcal{B}_T(\theta_\ell^\star, r)$ ("near region"). Thus, it is enough to prove that $|\eta_{\alpha,\xi}(\theta)| \leq 1 - C_N\,\mathfrak{d}_T(\theta_\ell^\star,\theta)^2$. This will be done by using Lemma 2.11.4 to obtain a quadratic decay on $\eta_{\alpha,\xi}$ from a bound on its second Riemannian derivative.

Recall that the function $\eta_{\alpha,\xi}$ is twice continuously differentiable. Set $\eta_{\alpha,\xi}^{[2]} = \tilde{D}_{2;T}[\eta_{\alpha,\xi}]$. Differentiating (2.87) and using that $\mathcal{K}_T^{[2,0]}(\theta,\theta) = -1$ and $\mathcal{K}_T^{[2,1]}(\theta,\theta) = 0$, see (2.29), we deduce that:

$$\eta_{\alpha,\xi}^{[2]}(\theta) = e_\ell^\top(I + \Gamma_{\ell,\theta}^{[2,0]})\alpha + \mathcal{K}_T^{[2,0]}(\theta,\theta_\ell^\star)e_\ell^\top\alpha + e_\ell^\top\Gamma_{\ell,\theta}^{[2,1]}\xi + \mathcal{K}_T^{[2,1]}(\theta,\theta_\ell^\star)e_\ell^\top\xi. \tag{2.94}$$

Since $v = (v_1,\cdots,v_s)^\top \in \{-1,1\}^s$ is a sign vector, we get:

$$\eta_{\alpha,\xi}^{[2]}(\theta) - v_\ell\mathcal{K}_T^{[2,0]}(\theta,\theta_\ell^\star) = e_\ell^\top(I+\Gamma_{\ell,\theta}^{[2,0]})\alpha + \mathcal{K}_T^{[2,0]}(\theta,\theta_\ell^\star)e_\ell^\top(\alpha - v) + e_\ell^\top\Gamma_{\ell,\theta}^{[2,1]}\xi + \mathcal{K}_T^{[2,1]}(\theta,\theta_\ell^\star)e_\ell^\top\xi. \tag{2.95}$$

The triangle inequality and the definition of $\mathcal{V}_T$ give:

$$|\mathcal{K}_T^{[2,0]}(\theta,\theta_\ell^\star)| \leq L_{2,0} + \mathcal{V}_T \quad \text{and} \quad |\mathcal{K}_T^{[2,1]}(\theta,\theta_\ell^\star)| \leq L_{2,1} + \mathcal{V}_T, \tag{2.96}$$

where $L_{2,0}$ and $L_{1,2}$ are defined in (2.31). We deduce from (2.91), the definition of δ_T in (2.39) and (2.40) that:

$$\left\|I + \Gamma_{\ell,\theta}^{[2,0]}\right\|_{\mathrm{op}} \leq u_T(s) \quad \text{and} \quad \left\|\Gamma_{\ell,\theta}^{[2,1]}\right\|_{\mathrm{op}} \leq u_T(s). \tag{2.97}$$

We deduce from (2.95) that:

$$\begin{aligned}
|\eta_{\alpha,\xi}^{[2]}(\theta) - v_\ell\mathcal{K}_T^{[2,0]}(\theta,\theta_\ell^\star)| &\leq \|\alpha\|_{\ell_\infty}\left\|I + \Gamma_{\ell,\theta}^{[2,0]}\right\|_{\mathrm{op}} + \|\alpha - v\|_{\ell_\infty}(L_{2,0} + \mathcal{V}_T) \\
&\quad + \|\xi\|_{\ell_\infty}\left(\left\|\Gamma_{\ell,\theta}^{[2,1]}\right\|_{\mathrm{op}} + L_{2,1} + \mathcal{V}_T\right) \\
&\leq \frac{u_T(s)}{1-2u_T(s)}(1 + L_{2,0} + L_{2,1} + 2\mathcal{V}_T).
\end{aligned}$$

By assumption, we have $u_T(s) \leq H_\infty^{(2)}(r,\rho) \leq 1/4$. Hence, we have $\frac{1}{1-2u_T(s)} \leq 2$. Furthermore, we have by assumption $\mathcal{V}_T \leq H_\infty^{(1)}(r,\rho) \leq 1/2$ and $u_T(s) \leq H_\infty^{(2)}(r,\rho)$. In particular, we have:

$$u_T(s) \leq \frac{8}{9(2L_{2,0} + 2L_{2,1} + 4)}\nu_\infty(\rho r).$$

Therefore, we obtain:

$$|\eta_{\alpha,\xi}^{[2]}(\theta) - v_\ell\mathcal{K}_T^{[2,0]}(\theta,\theta_\ell^\star)| \leq \frac{8}{9}\nu_\infty(\rho r). \tag{2.98}$$

We now check that the hypotheses of Lemma 2.11.4-*(ii)* hold in order to obtain a quadratic decay on $\eta_{\alpha,\xi}$ from the bound (2.98). First recall that $\eta_{\alpha,\xi}$ is twice continuously differentiable and have the interpolation properties (2.88). By the triangle inequality and since by assumption $\mathcal{V}_T \leq L_{2,0}$ we have:

$$\sup_{\Theta_T^2}|\mathcal{K}_T^{[2,0]}| \leq L_{2,0} + \mathcal{V}_T \leq 2L_{2,0}.$$

Then, Lemma 2.7.1 ensures that for any θ, θ' in Θ_T such that $\mathfrak{d}_T(\theta, \theta') \leq r$ we have:

$$-\mathcal{K}_T^{[2,0]}(\theta, \theta') \geq \nu_\infty(r\rho_T) - \mathcal{V}_T \geq \nu_\infty(\rho r) - \mathcal{V}_T \geq \frac{9}{10}\nu_\infty(\rho r),$$

where we used that that the function $r \mapsto \nu_\infty(r)$ is decreasing and $\rho_T \leq \rho$ for the second inequality and that $\mathcal{V}_T \leq H_\infty^{(1)}(r, \rho) \leq \nu_\infty(\rho r)/10$ for the last inequality.

Set $\delta = \frac{8}{9}\nu_\infty(\rho r)$, $\varepsilon = \frac{9}{10}\nu_\infty(\rho r)$, $L = 2L_{2,0}$. As $r < L^{-\frac{1}{2}}$ and $\delta < \varepsilon$, we apply Lemma 2.11.4-*(ii)* and get for $\theta \in \mathcal{B}_T(\theta_\ell^\star, r)$:

$$|\eta_{\alpha,\xi}(\theta)| \leq 1 - \frac{\nu_\infty(\rho r)}{180}\, \mathfrak{d}_T(\theta, \theta_\ell^\star)^2.$$

Proof of *(ii)* from Assumption 2.6.1 with $C'_N = (5L_{2,0} + L_{2,1} + 4)/8$. Let $\theta \in \Theta_T$ such that $\mathfrak{d}_T(\theta, \mathcal{Q}^\star) \leq r$. Let $\ell \in \{1, \cdots, s\}$ such that $\theta \in \mathcal{B}_T(\theta_\ell^\star, r)$ ("near region"). We shall prove that $|\eta_{\alpha,\xi}(\theta) - v_\ell| \leq C'_N\, \mathfrak{d}_T(\theta_\ell^\star, \theta)^2$.

Let us consider the function $f : \theta \to \eta_{\alpha,\xi}(\theta) - v_\ell$. We will bound the second covariant derivative $f^{[2]} = \tilde{D}_{2;T}[f]$ of f and apply Lemma 2.11.4-*(i)* on f to prove the property *(ii)* for $\eta_{\alpha,\xi}$. Notice that f is twice continuously differentiable. By construction, see (2.88), we have $f(\theta_\ell^\star) = 0$ and $f^{[1]}(\theta_\ell^\star) = 0$. Since $f^{[2]} = \eta_{\alpha,\xi}^{[2]}$, we deduce from (2.94), the bounds (2.96) that:

$$|f^{[2]}(\theta)| \leq \|\alpha\|_{\ell_\infty} \left\| I + \Gamma_{\ell,\theta}^{[2,0]} \right\|_{\text{op}} + \|\alpha\|_{\ell_\infty}(L_{2,0} + \mathcal{V}_T) + \|\xi\|_{\ell_\infty} \left\| \Gamma_{\ell,\theta}^{[2,1]} \right\|_{\text{op}} + \|\xi\|_{\ell_\infty}(L_{2,1} + \mathcal{V}_T).$$

Using (2.97), and the bounds on α and ξ from Lemma 2.10.1, we get:

$$|f^{[2]}(\theta)| \leq \frac{1 - u_T(s)}{1 - 2u_T(s)}(L_{2,0} + \mathcal{V}_T + u_T(s)) + \frac{u_T(s)}{1 - 2u_T(s)}(L_{2,1} + \mathcal{V}_T + u_T(s)).$$

Since $u_T(s) \leq H_\infty^{(2)}(r, \rho) \leq 1/6$ and $\mathcal{V}_T \leq H_\infty^{(1)}(r, \rho) \leq 1/2$, we get:

$$|f^{[2]}(\theta)| \leq \frac{5}{4}L_{2,0} + \frac{1}{4}L_{2,1} + 1.$$

We get thanks to Lemma 2.11.4-*(i)* on the function f that for any $\theta \in \mathcal{B}_T(\theta_\ell^\star, r)$:

$$|\eta_{\alpha,\xi}(\theta) - v_\ell| \leq \frac{1}{8}\,(5L_{2,0} + L_{1,2} + 4)\, \mathfrak{d}_T(\theta, \theta_\ell^\star)^2.$$

Proof of *(iv)* from Assumption 2.6.1 with $C_B = 2$. Recall the definition of $p_{\alpha,\xi}$ in (2.85). Elementary calculations give using the definitions of $\Gamma^{[0,0]}$, $\Gamma^{[1,1]}$ and $\Gamma^{[1,1]}$ in (2.83):

$$\begin{aligned}
\|p_{\alpha,\xi}\|_T^2 &\leq 2\left\| \sum_{k=1}^s \alpha_k \phi_T(\theta_k^\star) \right\|_T^2 + 2\left\| \sum_{k=1}^s \xi_k\, \phi_T^{[1]}(\theta_k^\star) \right\|_T^2 \\
&= 2\alpha^\top \Gamma^{[0,0]} \alpha + 2\xi^\top \Gamma^{[1,1]} \xi \\
&\leq 2\|\alpha\|_{\ell_1}\|\alpha\|_{\ell_\infty} \left\| \Gamma^{[0,0]} \right\|_{\text{op}} + 2\|\xi\|_{\ell_1}\|\xi\|_{\ell_\infty} \left\| \Gamma^{[1,1]} \right\|_{\text{op}}.
\end{aligned}$$

Using that $\|I\|_{\text{op}} = 1$ and (2.84), we get that:

$$\left\| \Gamma^{[0,0]} \right\|_{\text{op}} \leq (1 + u_T(s)) \quad \text{and} \quad \left\| \Gamma^{[1,1]} \right\|_{\text{op}} \leq (1 + u_T(s)).$$

By assumption we have $u_T(s) \leq H_\infty^{(2)}(r, \rho) \leq \frac{1}{6}$. We deduce that:

$$\|p_{\alpha,\xi}\|_T^2 \leq 2(1 + u_T(s))\frac{(1 - u_T(s))^2 + u_T(s)^2}{(1 - 2u_T(s))^2} s \leq 4s.$$

This gives:

$$\|p_{\alpha,\xi}\|_T \leq 2\sqrt{s}. \tag{2.99}$$

We proved that *(i)*-*(iv)* from Assumption 2.6.1 stand. By assumption we also have that for all $\theta \neq \theta' \in \mathcal{Q}^\star : \mathfrak{d}_T(\theta, \theta') > 2\,r$, therefore Assumption 2.6.1 holds.

This finishes the proof of Proposition 2.7.4.

2.10.2 Proof of Proposition 2.7.5 (Construction of an interpolating derivative certificate)

This section is devoted to the proof of Proposition 2.7.5 and is close to Section 2.10.1. Let $T \in \mathbb{N}$ and $s \in \mathbb{N}^*$. Recall Assumptions 2.3.2 (and thus 2.3.1 on the regularity of φ_T) and 2.5.1 on the regularity of the limit kernel $\mathcal{K}_\infty$ are in force. Set $u'_\infty \in (0, 1/6)$. We denote by $\|\cdot\|_{\mathrm{op}}$ the operator norm associated to the ℓ_∞ norm on $\mathbb{R}^s$. By assumption $\delta_\infty(u'_\infty, s)$ is finite. Let $\vartheta^\star = (\theta^\star_1, \dots, \theta^\star_s) \in \Theta^s_{T, 2\rho_T\, \delta_\infty(u'_\infty, s)}$. We note $\mathcal{Q}^\star = \{\theta^\star_i,\, 1 \leq i \leq s\}$ the set of parameters of cardinal s. Let $\alpha = (\alpha_1, \dots, \alpha_s)^\top$ and $\xi = (\xi_1, \dots, \xi_s)^\top$ be elements of $\mathbb{R}^s$. Recall $p_{\alpha,\xi}$, $\eta_{\alpha,\xi}$ and $\eta^{[1]}_{\alpha,\xi} = \tilde{D}_{1;T}[\eta_{\alpha,\xi}]$ given by (2.85), (2.86) and (2.87).

The next lemma is similar to Lemma 2.10.1, but notice that in Lemma 2.10.2 the function $\eta_{\alpha,\xi}$ vanished on $\mathcal{Q}^\star$ and has a derivative that interpolates a sign vector, whereas in Lemma 2.10.1 it is the opposite.

Recall the definition of $\mathcal{V}_T$ from (2.34) and define $u'_T(s) = u'_\infty + (s-1)\mathcal{V}_T$. We remark that (2.84) holds with $u_T(s)$ replaced by $u'_T(s)$ because of (2.82).

Lemma 2.10.2. *Let $v = (v_1, \cdots, v_s)^\top \in \{-1, 1\}^s$ be a sign vector. Assume that (2.84) holds with $u_T(s)$ replaced by $u'_T(s) < 1/2$. Under Assumption 2.3.2, there exist unique $\alpha, \xi \in \mathbb{R}^s$ such that:*

$$\eta_{\alpha,\xi}(\theta^\star_k) = 0 \quad \text{and} \quad \eta^{[1]}_{\alpha,\xi}(\theta^\star_k) = v_k \quad \text{for} \quad 1 \leq k \leq s. \tag{2.100}$$

Furthermore, we have:

$$\|\alpha\|_{\ell_\infty} \leq \frac{u'_T(s)}{1 - 2u'_T(s)} \quad \text{and} \quad \|\xi\|_{\ell_\infty} \leq \frac{1 - u'_T(s)}{1 - 2u'_T(s)}. \tag{2.101}$$

Proof. Thus, with 0_s the vector of size s with all its components equal to zero and Γ defined by (2.83), Equation (2.100) is equivalent to:

$$\Gamma \begin{pmatrix} \alpha \\ \xi \end{pmatrix} = \begin{pmatrix} 0_s \\ v \end{pmatrix}. \tag{2.102}$$

According to the proof of Lemma 2.10.1, the matrices $\Gamma_{SC} = \Gamma^{[0,0]} - \Gamma^{[1,0]\top}[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}$, $\Gamma^{[1,1]}$ and Γ are non singular. Thus the vectors α and ξ exist and are uniquely determined by (2.102). From Lemma 2.11.3, we deduce that:

$$\alpha = -\Gamma_{SC}^{-1}\Gamma^{[1,0]\top}[\Gamma^{[1,1]}]^{-1}v \quad \text{and} \quad \xi = \left(I + [\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}\Gamma_{SC}^{-1}\Gamma^{[1,0]\top}\right)[\Gamma^{[1,1]}]^{-1}v.$$

Using (2.90), (2.84) and (2.89) and replacing $u_T(s)$ by $u'_T(s)$, we easily obtain the inequalities (2.101). □

We fix the sign vector $v = (v_1, \cdots, v_s)^\top \in \{-1, 1\}^s$ and consider $p_{\alpha,\xi}$ and $\eta_{\alpha,\xi}$ given by (2.85) and (2.86), with α and ξ given by Lemma 2.10.2.

Proof of (*i*) from Assumption 2.6.2 with $c_N = (L_{0,2} + L_{2,1} + 7)/8$. We define the function $f : \theta \mapsto \eta_{\alpha,\xi}(\theta) - v_\ell\, \mathrm{sign}(\theta - \theta^\star_\ell)\mathfrak{d}_T(\theta, \theta^\star_\ell)$ on Θ. To prove the Property (*i*), we will bound the second covariant derivative of f, that is $f^{[2]} := \tilde{D}_{2;T}[f]$, and apply Lemma 2.11.4-(*i*). Recall $\mathfrak{d}_T(\theta, \theta^\star_\ell) = |G_T(\theta) - G_T(\theta^\star_\ell)|$ with G_T a primitive of $\sqrt{g_T}$, and thus $f(\theta) = \eta_{\alpha,\xi}(\theta) - v_\ell(G_T(\theta) - G_T(\theta^\star_\ell))$. We deduce that f is twice continuously differentiable on Θ; and elementary calculations give $f^{[2]} = \eta^{[2]}_{\alpha,\xi}$.

Let $\theta \in \Theta_T$ and let $\theta^\star_\ell$ be one of the elements of $\mathcal{Q}^\star$ closest to θ in terms of the metric $\mathfrak{d}_T$. Recall the notations $\Gamma_{\ell,\theta}$ (resp. $\Gamma^{[i,j]}_{\ell,\theta}$) and $\vartheta^\star_{\ell,\theta}$ from the proof of Proposition 2.7.4. Since $f^{[2]} = \eta^{[2]}_{\alpha,\xi}$, we deduce from (2.94) that:

$$|f^{[2]}(\theta)| \leq \left\|I + \Gamma^{[2,0]}_{\ell,\theta}\right\|_{\mathrm{op}} \|\alpha\|_{\ell_\infty} + \|\alpha\|_{\ell_\infty} |\mathcal{K}^{[2,0]}_T(\theta, \theta^\star_\ell)| + \|\xi\|_{\ell_\infty} \left\|\Gamma^{[2,1]}_{\ell,\theta}\right\|_{\mathrm{op}} + \|\xi\|_{\ell_\infty} |\mathcal{K}^{[2,1]}_T(\theta, \theta^\star_\ell)|.$$

Notice that (2.91) holds with $u_T(s)$ replaced by $u'_T(s)$. Using (2.96) and (2.97) and the bounds (2.101) on α and ξ from Lemma 2.10.2, we get:

$$|f^{[2]}(\theta)| \leq \frac{u'_T(s)}{1-2u'_T(s)}(L_{2,0} + \mathcal{V}_T + u'_T(s)) + \frac{1-u'_T(s)}{1-2u'_T(s)}(L_{2,1} + \mathcal{V}_T + u'_T(s)).$$

By assumption, we have $u'_T(s) \leq 1/6$ and $\mathcal{V}_T \leq 1$. Hence, we obtain:

$$|f^{[2]}(\theta)| \leq \frac{1}{4}L_{2,0} + \frac{5}{4}L_{2,1} + \frac{7}{4}.$$

Since $f(\theta^\star_\ell) = 0$ and $f^{[1]}(\theta^\star_\ell) = 0$ as well, using Lemma 2.11.4 *(i)*, we get, with $c_N = (L_{2,0} + 5L_{2,1} + 7)/8$:

$$|\eta_{\alpha,\xi}(\theta) - v_\ell \operatorname{sign}(\theta - \theta^\star_\ell)\, \mathfrak{d}_T(\theta, \theta^\star_\ell)| = |f(\theta)| \leq c_N\, \mathfrak{d}_T(\theta, \theta^\star_\ell)^2.$$

Proof of *(ii)* from Assumption 2.6.2 with $c_F = (5L_{1,0} + 7)/4$. Let $\theta \in \Theta_T$, we shall prove that $|\eta_{\alpha,\xi}(\theta)| \leq c_F$. Let $\theta^\star_\ell$ be one of the elements of $\mathcal{Q}^\star$ closest to θ in terms of the metric $\mathfrak{d}_T$. We deduce from (2.92) that:

$$|\eta_{\alpha,\xi}(\theta)| \leq \|\alpha\|_{\ell_\infty} \left\|\Gamma^{[0,0]}_{\ell,\theta} - I\right\|_{\text{op}} + \|\alpha\|_{\ell_\infty} |\mathcal{K}_T(\theta, \theta^\star_\ell)| + \|\xi\|_{\ell_\infty} \left\|\Gamma^{[1,0]\top}_{\ell,\theta}\right\|_{\text{op}} + \|\xi\|_{\ell_\infty} |\mathcal{K}^{[0,1]}_T(\theta, \theta^\star_\ell)|.$$

Using (2.84), (2.29), (2.93) and the bounds (2.101) on α and ξ from Lemma 2.10.2, we get:

$$|\eta_{\alpha,\xi}(\theta)| \leq \frac{u'_T(s)}{1-2u'_T(s)}\left(1 + u'_T(s)\right) + \frac{1-u'_T(s)}{1-2u'_T(s)}\left(L_{1,0} + \mathcal{V}_T + u'_T(s)\right).$$

By assumption, we have $u'_T(s) \leq 1/6$, and thus $\frac{1}{1-2u'_T(s)} \leq 3/2$. Since $\mathcal{V}_T \leq 1$, we obtain:

$$|\eta_{\alpha,\xi}(\theta)| \leq \frac{5}{4}L_{1,0} + \frac{7}{4}.$$

Proof of *(iii)* from Assumption 2.6.2 with $c_B = 2$. Using very similar arguments as in the proof of (2.99) (taking care that the upper bound of the ℓ_∞ norm of α and ξ are given by (2.101)) we also get $\|p_{\alpha,\xi}\|_T \leq 2\sqrt{s}$.

We proved that *(i)*-*(ii)* from Assumption 2.6.2 stand for any $\theta \in \Theta_T$. Hence Assumption 2.6.2 holds for any positive r such that for all $\theta \neq \theta' \in \mathcal{Q}^\star : \mathfrak{d}_T(\theta, \theta') > 2\,r$.

This finishes the proof of Proposition 2.7.5.

2.11 Auxiliary Lemmas

In this section we provide the proofs of the intermediate results of Chapter 2.

2.11.1 Tail bounds for suprema of Gaussian processes

In order to prove Theorems 2.2.1 and 2.2.5, we provide in Lemma 2.11.1 a bound with high probability of the supremum of a Gaussian process given by $\theta \mapsto \langle w_T, h(\theta)\rangle_T$, where w_T is a noise process and h is a function from Θ, an interval of $\mathbb{R}$, to the Hilbert space $(H_T, \langle\cdot,\cdot\rangle_T)$. The next lemma is in the spirit of [Azaïs and Wschebor, 2009, Proposition 4.1] (where one assumes that the Gaussian process has unitary variance); its proof is given at the end of this section and relies on Lemma 2.11.2.

We denote by $\mathfrak{d}_T$ the Riemannian metric associated to the kernel $\mathcal{K}_T$, see also Section 2.4.2. Recall definitions (2.17) and (2.19) and set $f^{[1]}(\theta) = \tilde{D}_{1,T}[f](\theta) = \partial_\theta f(\theta)/\sqrt{g_T(\theta)}$ with g_T defined in (2.27).

Lemma 2.11.1. *Let $T \in \mathbb{N}$ be fixed. Suppose that Assumptions 2.3.1 and 2.3.2 hold. Let h be a function of class $\mathcal{C}^1$ from Θ_T to H_T, with Θ_T a sub-interval of Θ. Assume there exist finite constants C_1 and C_2 such that for all $\theta \in \Theta_T$:*

$$\|h(\theta)\|_T \leq C_1 \quad \text{and} \quad \left\|h^{[1]}(\theta)\right\|_T \leq C_2.$$

Let w_T be an H_T-valued Gaussian noise such that Assumption 2.1.1 holds, and consider the Gaussian process $X = (X(\theta) = \langle h(\theta), w_T\rangle_T, \theta \in \Theta)$. Then, we have for $u > 0$:

$$\mathbb{P}\left(\sup_{\theta\in\Theta_T} |X(\theta)| \geq u\right) \leq c \cdot \left(\sigma\frac{|\Theta_T|\sqrt{\Delta_T}}{u} \vee 1\right) \mathrm{e}^{-u^2/(4\sigma^2\Delta_T C_1^2)}, \tag{2.103}$$

where $|\Theta_T|$ denotes the Riemannian length of the interval Θ_T and $c = 2C_2 + 1$.

We first state a technical lemma.

Lemma 2.11.2. *Let $I \subset \mathbb{R}$ be an interval. Assume that $X = (X(\theta),\, \theta \in I)$ is a real centered Gaussian process with Lipschitz sample paths. Then, for all $u > 0$ and an arbitrary $\theta_0 \in I$, we have:*

$$\mathbb{P}\left(\sup_I X \geq u\right) \leq \frac{1}{u}\int_I \sqrt{\mathrm{Var}(X'(\theta))}\, \mathrm{e}^{-u^2/(4\mathrm{Var}(X(\theta)))}\, \mathrm{d}\theta + \frac{1}{2}\mathrm{e}^{-u^2/(2\mathrm{Var}(X(\theta_0)))}. \tag{2.104}$$

Proof. We first start with a general remark on Lipschitz functions on $\mathbb{R}$. Let f be a real-valued Lipschitz function defined on an interval $I \subset \mathbb{R}$. Let $b > a$ and set $f_{a,b} = \min(\max(f, a), b)$. The function $f_{a,b}$ is also Lipschitz and, thanks to [Evans and Gariepy, 2015, Theorem 3.3 p107], we get that $f'_{a,b} = f' = 0$ a.e. on $\{x \in I :\ f(x) = a \text{ or } b\}$ and thus $f'_{a,b} = f'\,\mathbf{1}_{\{f\in(a,b)\}}$ a.e. on I. We deduce that:

$$\sup f_{a,b} - \inf f_{a,b} \leq \int_I |f'_{a,b}(x)|\,\mathrm{d}x = \int_I |f'(x)|\,\mathbf{1}_{\{f(x)\in(a,b)\}}\,\mathrm{d}x.$$

Using this inequality, we obtain that for any $x_0 \in I$:

$$\int_a^b \mathbf{1}_{\{\sup_I f>t\}}\,\mathrm{d}t = \int_a^b \mathbf{1}_{\{\sup_I f_{a,b}>t\}}\,\mathrm{d}t = \sup f_{a,b} - a \leq (b-a)\mathbf{1}_{\{f(x_0)\geq a\}} + \int_I |f'(x)|\,\mathbf{1}_{\{f(x)\in(a,b)\}}\,\mathrm{d}x. \tag{2.105}$$

Then, applying Inequality (2.105) to the function X and taking the expectation, we get, with $M = \sup_I X$, $a = u > 0$, $b = u + \varepsilon$, $\varepsilon > 0$ and $x_0 = \theta_0$:

$$\int_u^{u+\varepsilon} \mathbb{P}(M \geq t)\,\mathrm{d}t \leq \varepsilon\mathbb{P}(X(\theta_0) \geq u) + \int_I \mathbb{E}\left[|X'(\theta)|\mathbf{1}_{\{u<X(\theta)<u+\varepsilon\}}\right]\mathrm{d}\theta. \tag{2.106}$$

The random variable $X(\theta_0)$ is a centered Gaussian variable and therefore we have:

$$\mathbb{P}\left(X(\theta_0) \geq u\right) = \int_u^{+\infty} \frac{\mathrm{e}^{-x^2/(2\mathrm{Var}(X(\theta_0)))}}{\sqrt{2\pi\mathrm{Var}(X(\theta_0))}}\,\mathrm{d}x \leq \frac{1}{2}\mathrm{e}^{-u^2/2\mathrm{Var}(X(\theta_0))}, \tag{2.107}$$

where we used for the inequality that $\int_u^{+\infty}\mathrm{e}^{-t^2}\,\mathrm{d}t \leq \frac{\sqrt{\pi}}{2}\,\mathrm{e}^{-u^2}$ holds for $u > 0$, see [Abramowitz and Stegun, 1992, Formula 7.1.13]. Notice that (2.107) trivially holds if $\mathrm{Var}(X(\theta_0)) = 0$ as $u > 0$.

We now give a bound of the second term in the right hand-side of (2.106). Since (X', X) is also a Gaussian process, we can write:

$$X'(\theta) = \alpha_\theta X(\theta) + \beta_\theta G,$$

where G is a standard Gaussian random variable independent of $X(\theta)$ and:

$$\alpha_\theta = \frac{\mathbb{E}[X'(\theta)X(\theta)]}{\mathrm{Var}(X(\theta))} \quad \text{and} \quad \beta_\theta^2 = \mathrm{Var}(X'(\theta)) - \alpha_\theta^2 \mathrm{Var}(X(\theta)),$$

with the convention that $\alpha_\theta = 0$ if $\mathrm{Var}(X(\theta)) = 0$. We get $|X'(\theta)| \leq |\alpha_\theta X(\theta)| + |\beta_\theta|\,|G|$. Since G is independent of $X(\theta)$ and $u > 0$, we deduce that:

$$\mathbb{E}\left[|X'(\theta)|\mathbf{1}_{\{u<X(\theta)<u+\varepsilon\}}\right] \leq \left(|\alpha_\theta|(u+\varepsilon) + \sqrt{\frac{2}{\pi}}\,|\beta_\theta|\right)\mathbb{P}(u < X(\theta) < u+\varepsilon).$$

Letting ε goes to 0 in (2.106), using (2.107) the right continuity of the cdf of M and the monotonicity of the density $p_{X(\theta)}(u)$ of the law of $X(\theta)$, we deduce that:

$$\mathbb{P}(M \geq u) \leq \frac{1}{2}\,\mathrm{e}^{-u^2/2\mathrm{Var}(X(\theta_0))} + \int_I \left(|\alpha_\theta|u + \sqrt{\frac{2}{\pi}}\,|\beta_\theta|\right) p_{X(\theta)}(u)\,\mathrm{d}\theta, \tag{2.108}$$

where by convention $p_{X(\theta)}(u)$ is taken equal to 0 if $\mathrm{Var}(X(\theta)) = 0$. We now bound the second term of the right-hand side of (2.108) in two steps. Using that $\beta_\theta^2 \leq \mathrm{Var}(X'(\theta))$ and the inequality $\mathrm{e}^{-x^2} \leq \mathrm{e}^{-x^2/2}/\sqrt{2}\,x$ for $x > 0$, we get that:

$$\sqrt{\frac{2}{\pi}}\,|\beta_\theta|\,p_{X(\theta)}(u) \leq \frac{1}{\pi}\,\frac{\sqrt{\mathrm{Var}(X'(\theta))}}{u}\,\mathrm{e}^{-u^2/4\mathrm{Var}(X(\theta))}. \tag{2.109}$$

Thanks to the Cauchy-Schwarz inequality, we get $|\alpha_\theta| \leq \sqrt{\mathrm{Var}(X'(\theta))}/\sqrt{\mathrm{Var}(X(\theta))}$. Using also the inequality $\mathrm{e}^{-x^2} \leq 3\,\mathrm{e}^{-x^2/2}/4x^2$ for $x > 0$, we get that:

$$|\alpha_\theta|u\,p_{X(\theta)}(u) \leq \frac{3}{4}\sqrt{\frac{2}{\pi}}\,\frac{\sqrt{\mathrm{Var}(X'(\theta))}}{u}\,\mathrm{e}^{-u^2/4\mathrm{Var}(X(\theta))}. \tag{2.110}$$

Notice that (2.109) and (2.110) hold also if $\mathrm{Var}(X(\theta)) = 0$. Using that $\frac{3}{4}\sqrt{\frac{2}{\pi}} + \frac{1}{\pi} \simeq 0.92 \leq 1$, we deduce (2.104) from (2.108), (2.109) and (2.110). □

Proof of Lemma 2.11.1. We first consider the case $\Theta_T = [\theta_0, \theta_1]$ and let $\gamma : [0,1] \to [\theta_0, \theta_1]$ be a minimizing path with respect to the Riemannian metric $\mathfrak{d}_T$ (see Remark 2.4.1); in particular we have $|\gamma'(s)|\sqrt{g_T(\gamma(s))} = \mathfrak{d}_T(\theta_0, \theta_1)$. Thanks to (2.9), the Gaussian process $\tilde{X} = (\tilde{X}(s) = X(\gamma(s)), s \in [0,1])$ is of class $\mathcal{C}^1$ on $s \in [0,1]$, with derivative $\tilde{X}'(s) = \gamma'(s)\,X'(\gamma(s)) = \gamma'(s)\,\langle \partial_\theta h(\gamma(s)), w_T\rangle_T$. Then, according to Lemma 2.11.2, Inequality (2.104) holds. By Assumption 2.1.1, we have for all $\theta \in \Theta_T$:

$$\mathrm{Var}(X(\theta))) \leq \sigma^2\Delta_T\|h(\theta)\|_T^2 \leq \sigma^2\Delta_T C_1^2 \quad \text{and} \quad \frac{\mathrm{Var}(X'(\theta)))}{g_T(\theta)} \leq \sigma^2\Delta_T\left\|h^{[1]}(\theta)\right\|_T^2 \leq \sigma^2\Delta_T C_2^2.$$

Plugging those bounds in Inequality (2.104) with $|\gamma'(s)|\sqrt{g_T(\gamma(s))} = \mathfrak{d}_T(\theta_0, \theta_1)$, we obtain:

$$\begin{aligned}\mathbb{P}\left(\sup_{[\theta_0,\theta_1]} X \geq u\right) &\leq \frac{1}{u}\sqrt{\sigma^2\Delta_T}C_2\,\mathrm{e}^{-u^2/(4\sigma^2\Delta_T C_1^2)}\int_0^1 |\gamma'(s)|\sqrt{g_T(\gamma(s))}\,\mathrm{d}s + \frac{1}{2}\,\mathrm{e}^{-u^2/(2\sigma^2\Delta_T C_1^2)} \\ &\leq \left(C_2 + \frac{1}{2}\right)\left(\sigma\frac{\mathfrak{d}_T(\theta_0,\theta_1)\sqrt{\Delta_T}}{u} \vee 1\right)\mathrm{e}^{-u^2/(4\sigma^2\Delta_T C_1^2)}.\end{aligned}$$

Since $\mathbb{P}\left(\sup_{[\theta_0,\theta_1]}|X| \geq u\right) \leq 2\,\mathbb{P}\left(\sup_{[\theta_0,\theta_1]} X \geq u\right)$, we obtain that (2.103) holds for Θ_T a bounded closed interval. Then, use monotone convergence and the continuity of X to get (2.103) for any interval Θ_T. □

2.11.2 Schur complement

The following Lemma is a classical result on the Schur complement.

Lemma 2.11.3 (Schur complement)**.** *Let $M \in \mathbb{R}^{n\times n}$ be a matrix composed of blocks $A \in \mathbb{R}^{(n-k)\times(n-k)}$, $B \in \mathbb{R}^{(n-k)\times k}$, $C \in \mathbb{R}^{k\times(n-k)}$, $D \in \mathbb{R}^{k\times k}$:*

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

Assume that D and $S_1 = A - BD^{-1}C$ are non singular. Then, the system:

$$M \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}.$$

with $x \in \mathbb{R}^{n-k}$, $y \in \mathbb{R}^k$, $a \in \mathbb{R}^{n-k}$ and $b \in \mathbb{R}^k$, has a unique solution given by:

$$x = S_1^{-1}a - S_1^{-1}BD^{-1}b \quad \text{and} \quad y = D^{-1}b - D^{-1}CS_1^{-1}a + D^{-1}CS_1^{-1}BD^{-1}b.$$

2.11.2.1 Proofs of Lemmas in Section 2.4

Proof of Lemma 2.4.2. For simplicity, we remove the subscript $\mathcal{K}$ and for example write $f^{[1]} = \tilde{D}_1[f] = D_1[f]/\sqrt{g}$. Recall that G, a primitive of $\sqrt{g}$, is continuous increasing and thus induces a one-to-one map from Θ to its image. Following Remark 2.4.1, we consider the minimizing path $\gamma : [0,1] \to \Theta$ from θ_0 to θ defined by $\gamma_s = G^{-1}(as+b)$, with $b = G(\theta_0)$ and $a = G(\theta) - G(\theta_0)$. Thus, we have $\mathcal{L}(\gamma) = \mathfrak{d}(\theta, \theta_0)$. The minimizing path from θ_0 to θ has constant speed thus equal to $\mathfrak{d}(\theta_0, \theta)$. From the explicit expression of γ, we get in fact that $\dot{\gamma}_t\sqrt{g(\gamma_t)} = A$ for $t \in [0,1]$, where $A = \operatorname{sign}(\theta - \theta_0)\,\mathfrak{d}(\theta, \theta_0)$. Thus, we have:

$$f(\theta) - f(\theta_0) = f(\gamma_1) - f(\gamma_0) = \int_0^1 \dot{\gamma}_t\, f'(\gamma_t)\, \mathrm{d}t = A\int_0^1 \tilde{D}_1[f](\gamma_t)\, \mathrm{d}t = A\int_0^1 f^{[1]}(\gamma_t)\, \mathrm{d}t, \tag{2.111}$$

where we used (2.10) and that the derivative of $f \circ \gamma_t$ is $\dot{\gamma}_t\, f' \circ \gamma_t$ for the second equality and the definition of $\tilde{D}_1[f]$ as well as the equality $\dot{\gamma}_t\sqrt{g(\gamma_t)} = A$ for the last.

Using (2.111) for f and θ replaced by $f^{[1]}$ and $\gamma(t)$ for some $t \in [0,1]$, we get thanks to (2.20) that:

$$f^{[1]}(\gamma_t) = f^{[1]}(\theta_0) + \tilde{A}\int_0^1 f^{[2]}(\tilde{\gamma}_s)\, \mathrm{d}s,$$

where $\tilde{\gamma}$ is a geodesic from θ_0 to γ_t and $\tilde{A} = \dot{\tilde{\gamma}}_s\sqrt{g(\tilde{\gamma}_s)}$. Since γ is itself a geodesic, we deduce that $\tilde{\gamma}_s = \gamma_{st}$, and thus $\tilde{A} = tA$. Plugging this in (2.111), we get:

$$f(\theta) - f(\theta_0) = A\, f^{[1]}(\theta_0) + A^2\int_{[0,1]^2} f^{[2]}(\gamma_{st})\; t\, \mathrm{d}t\, \mathrm{d}s = A\, f^{[1]}(\theta_0) + A^2\int_0^1 (1-r)\, f^{[2]}(\gamma_r)\, \mathrm{d}r.$$

This gives (2.21). □

Proof of Lemma 2.4.3. Recall that by Assumption 2.3.2 the function ϕ_T is $\mathcal{C}^3$. According to (2.9), we have that for any $i, j \in \{0, \dots, 3\}$ and any $\theta, \theta' \in \Theta$:

$$\partial^{i,j}_{\theta,\theta'} \left\langle \phi_T(\theta), \phi_T(\theta') \right\rangle_T = \left\langle \partial^i_\theta \phi_T(\theta), \partial^j_{\theta'} \phi_T(\theta') \right\rangle_T.$$

This and (2.17), (2.19), (2.22) and (2.23) readily imply (2.28). The first equality of (2.29) comes from Cauchy-Schwarz's inequality. The second is clear. We also have:

$$\langle \partial_\theta \phi_T(\theta), \phi_T(\theta) \rangle_T = \frac{1}{2}\, \partial_\theta \|\phi_T(\theta)\|^2 = 0 \tag{2.112}$$

Since the right hand-side is also equal to $\sqrt{g_T(\theta)}\,\mathcal{K}_T^{[1,0]}(\theta,\theta)$ thanks to (2.28), we get the third equality of (2.29). Taking the derivative with respect to θ in (2.112) yields $g_T(\theta) = \langle \partial_\theta \phi_T(\theta), \partial_\theta \phi_T(\theta)\rangle = -\langle \partial_\theta^2 \phi_T(\theta), \phi_T(\theta)\rangle$. Thanks to (2.18), we get $\partial_\theta^2 \phi_T = g_T \tilde{D}_{2,T}[\phi_T] + (1/2g_T) g_T' \partial_\theta \phi_T$. Using (2.28) and (2.112) again, we deduce that:

$$\langle \partial_\theta^2 \phi_T(\theta), \phi_T(\theta)\rangle = g_T(\theta)\,\mathcal{K}_T^{[2,0]}(\theta,\theta).$$

This gives the fourth equality of (2.29). Eventually, we deduce from (2.28), (2.18) and (2.19) that:

$$g_T(\theta)^{3/2}\,\mathcal{K}_T^{[2,1]}(\theta,\theta) = \langle \partial_\theta^2 \phi_T(\theta), \partial_\theta \phi_T(\theta)\rangle - \frac{1}{2}\frac{g_T'(\theta)}{g_T(\theta)}\langle \partial_\theta \phi_T(\theta), \partial_\theta \phi_T(\theta)\rangle.$$

Then, use that $g_T'(\theta) = 2\langle \partial_\theta^2 \phi_T(\theta), \partial_\theta \phi_T(\theta)\rangle$ to deduce that $\mathcal{K}_T^{[2,1]}(\theta,\theta) = 0$. □

2.11.3 Control on f from its derivatives $f^{[2]}$

The proof of the next lemma is similar to the proof of [Poon et al., 2021, Lemma 2] and is left to the reader. Recall from (2.29) that $\mathcal{K}_T^{[2,0]}(\theta,\theta) = -1$ on Θ.

Lemma 2.11.4. *Suppose Assumptions 2.3.1 and 2.3.2 on the dictionary hold. Let f be a real valued function defined on an interval Θ of class $\mathcal{C}^2$. Let $\theta_0 \in \Theta$. Set for $i = 1, 2$, $f^{[i]} = \tilde{D}_{i;T}[f]$ (see (2.19)).*

(i) Assume $f(\theta_0) = 0$, $f^{[1]}(\theta_0) = 0$ and that there exist $\delta > 0$ and $r > 0$ such that for any $\theta \in \mathcal{B}_T(\theta_0, r)$:

$$|f^{[2]}(\theta)| \le 2\delta.$$

Then, we have $|f(\theta)| \le \delta\, \mathfrak{d}_T(\theta,\theta_0)^2$, for any $\theta \in \mathcal{B}_T(\theta_0, r)$.

(ii) Let $\Theta_T \subset \Theta$ be an interval and suppose that $L \ge \sup_{\Theta_T^2} |\mathcal{K}_T^{[2,0]}|$ is finite and there exist $\varepsilon > 0$ and $r \in (0, L^{-\frac{1}{2}})$ such that for any $\theta \in \mathcal{B}_T(\theta_0, r)$, $-\mathcal{K}_T^{[2,0]}(\theta,\theta_0) \ge \varepsilon$. Assume that $\mathcal{B}_T(\theta_0, r) \subset \Theta_T$, $f(\theta_0) = v \in \{-1; 1\}$, $f^{[1]}(\theta_0) = 0$ and that there exists $\delta \in (0, \varepsilon)$ such that for any $\theta \in \mathcal{B}_T(\theta_0, r)$:

$$|f^{[2]}(\theta) - v\mathcal{K}_T^{[2,0]}(\theta,\theta_0)| \le \delta.$$

Then, we have $|f(\theta)| \le 1 - \frac{(\varepsilon-\delta)}{2}\mathfrak{d}_T(\theta,\theta_0)^2$, for any $\theta \in \mathcal{B}_T(\theta_0, r)$.

2.11.4 Proof of Lemma 2.8.1

We keep the notations from Section 2.8.1. In order to prove that the constants c_0, c_1 and c_2 do not depend on the scaling factor σ_0, we shall rewrite ρ_T and $\mathcal{V}_T$ defined in (2.32) and (2.34) using a change of scale. To do so, we define $\varphi^0(\theta) = k(\cdot - \theta)$ for $\theta \in \Theta$; the grid $t_1^0, \cdots, t_T^0$ where $t_j^0 = t_j/\sigma_0$; the Hilbert space $L^2(\lambda_T^0)$ with $\lambda_T^0 = \Delta_T\, \sigma_0^{-1} \sum_{j=1}^T \delta_{t_j^0}$, endowed with its natural scalar product noted $\langle \cdot, \cdot \rangle_{\lambda_T^0}$ and norm $\|\cdot\|_{\lambda_T^0}$; the parameter space $\Theta_T^0 = [a_T(1-\epsilon)\sigma_0^{-1}, b_T(1-\epsilon)\sigma_0^{-1}]$. Since the scaling factor σ_0 is fixed, the measures $(\lambda_T^0, T \ge 2)$ converge vaguely towards the Lebesgue measure λ_∞ on $\mathbb{R}$. We shall also consider another kernel:

$$\mathcal{K}_T^0(\theta,\theta') = \left\langle \phi_T^0(\theta), \phi_T^0(\theta') \right\rangle_{\lambda_T^0} \quad \text{with} \quad \phi_T^0 = \varphi^0 / \left\|\varphi^0\right\|_{\lambda_T^0},$$

and the limit kernel $\mathcal{K}_\infty^0(\theta,\theta') = \langle \phi_\infty^0(\theta), \phi_\infty^0(\theta')\rangle_\infty$ with $\phi_\infty^0 = \varphi^0/\|\varphi^0\|_\infty$. For any $T \in \mathbb{N} \cup \{+\infty\}$, the kernel $\mathcal{K}_T^0$ is of class $\mathcal{C}^{3,3}$ on Θ^2 and for $i, j \in \{0, \ldots, 3\}$ and $\theta, \theta' \in \Theta$, we have:

$$\mathcal{K}_T^{[i,j]}(\theta,\theta') = \mathcal{K}_T^{0[i,j]}\left(\frac{\theta}{\sigma_0}, \frac{\theta'}{\sigma_0}\right) \quad \text{and} \quad \frac{1}{\sigma_0^2} g_{\mathcal{K}_T^0}\left(\frac{\theta}{\sigma_0}\right) = g_{\mathcal{K}_T}(\theta).$$

We can now rewrite ρ_T and $\mathcal{V}_T$ by using a change of scale and we get:

$$\rho_T = \max\left(\sup_{\Theta_T^0}\sqrt{\frac{g_{\mathcal{K}_T^0}}{g_{\mathcal{K}_\infty^0}}}, \sup_{\Theta_T^0}\sqrt{\frac{g_{\mathcal{K}_\infty^0}}{g_{\mathcal{K}_T^0}}}\right),$$

and

$$\mathcal{V}_T = \max(\mathcal{V}_T^{(1)}, \mathcal{V}_T^{(2)}) \text{ with } \mathcal{V}_T^{(1)} = \max_{i,j\in\{0,1,2\}} \sup_{(\Theta_T^0)^2} |\mathcal{K}_T^{0[i,j]} - \mathcal{K}_\infty^{0[i,j]}| \text{ and } \mathcal{V}_T^{(2)} = \sup_{\Theta_T^0} |h_{\mathcal{K}_T^0} - h_{\mathcal{K}_\infty^0}|.$$

Thus, bounding ρ_T and $\mathcal{V}_T$ amounts to controling the proximity between the kernels $\mathcal{K}_T^0$ and $\mathcal{K}_\infty^0$.

First, we provide an upper bound for any $i, j \in \{0, \cdots, 3\}$ of:

$$B_{i,j}(T) = \sup_{\theta,\theta'\in\Theta_T^0} \left| \left\langle \partial_\theta^i \varphi^0(\theta), \partial_\theta^j \varphi^0(\theta') \right\rangle_{\lambda_T^0} - \left\langle \partial_\theta^i \varphi^0(\theta), \partial_\theta^j \varphi^0(\theta') \right\rangle_\infty \right|.$$

Notice that:

$$\partial_\theta^i \partial_t^j \varphi^0(\theta, t) = (-1)^j \, k^{(i+j)}(\theta - t).$$

Recall the polynomials P_i defined as $k^{(i)} = P_i \, k$. And set $M = \max_{0\le i\le 4} \sup |P_i| \sqrt{k}$. It is elementary to get that for $\theta, \theta' \in \mathbb{R}$:

$$\left| (\Delta_T/\sigma_0) \sum_{k=1}^{T} \partial_\theta^i \varphi^0(\theta, t_k^0) \partial_\theta^j \varphi^0(\theta', t_k^0) - \int_{a_T/\sigma_0}^{b_T/\sigma_0} \partial_\theta^i \varphi^0(\theta, t) \partial_\theta^j \varphi^0(\theta', t) \, \mathrm{d}t \right| \le 4\sqrt{\pi}\, \Delta_T M^2 \sigma_0^{-1}.$$

We have for $\theta, \theta' \in \Theta_T^0$ that:

$$\begin{aligned}
\left| \int_{\mathbb{R}\setminus[a_T/\sigma_0, b_T/\sigma_0]} \partial_\theta^i \varphi^0(\theta, t) \partial_\theta^j \varphi^0(\theta', t) \, \mathrm{d}t \right| &\le \left| \int_{b_T/\sigma_0}^{+\infty} \partial_\theta^i \varphi^0(\theta, t) \partial_\theta^j \varphi^0(\theta', t) \, \mathrm{d}t \right| \\
&\quad + \left| \int_{-\infty}^{a_T/\sigma_0} \partial_\theta^i \varphi^0(\theta, t) \partial_\theta^j \varphi^0(\theta', t) \, \mathrm{d}t \right| \\
&\le 2M^2 \int_{\epsilon b_T/\sigma_0}^{+\infty} k(t) \, \mathrm{d}t \\
&\le 2\sqrt{\pi}\, M^2 \, \mathrm{e}^{-\epsilon^2 b_T^2/2\sigma_0^2},
\end{aligned}$$

where we used that $2\int_u^{+\infty} \mathrm{e}^{-t^2} \, \mathrm{d}t \le \sqrt{\pi}\, \mathrm{e}^{-u^2}$ for $u > 0$, see formula 7.1.13 in [Abramowitz and Stegun, 1992]. We deduce that:

$$B_{i,j}(T) \le 4\sqrt{\pi}\, \Delta_T M^2 \sigma_0^{-1} + 2\sqrt{\pi}\, M^2 \, \mathrm{e}^{-\epsilon^2 b_T^2/2\sigma_0^2} \le 2\sqrt{\pi}\, M^2 \gamma_T,$$

with $\gamma_T = 2\Delta_T \sigma_0^{-1} + \sqrt{\pi}\, \mathrm{e}^{-\epsilon^2 b_T^2/2\sigma_0^2}$.

Similar arguments as above yield that:

$$\sup_{\theta\in\Theta_T^0} \left| \left\| \varphi^0(\theta) \right\|_{\lambda_T^0}^2 - \left\| \varphi^0(\theta) \right\|_\infty^2 \right| \le \gamma_T.$$

so that $\left\|\varphi^0(\theta)\right\|_{\lambda_T^0}^2 \ge \sqrt{\pi} - \gamma_T$ for all $\theta \in \Theta_T^0$. It is then easy to deduce that $\sup_{\Theta_T^0} |g_{\mathcal{K}_T^0} - g_{\mathcal{K}_\infty^0}|$ is bounded by a constant times γ_T when γ_T is smaller than a universal finite constant. Up to taking γ_T smaller than some universal finite constant, this and the fact that $g_{\mathcal{K}_\infty^0} = 1/2$ give the second part of (2.47). Then use formulae for the derivatives of the kernels, see (2.26) and (2.19), to get the first part of (2.47).

3 Simultaneous off-the-grid learning of mixtures issued from a continuous dictionary

Contents

Preamble

In this chapter we observe a set, possibly a continuum, of signals corrupted by noise. Each signal is a finite mixture of an unknown number of features belonging to a continuous dictionary. The continuous dictionary is parametrized by a real non-linear parameter. We shall assume that the signals share an underlying structure by saying that the union of active features in the whole dataset is finite.

We formulate regularized optimization problems to estimate simultaneously the linear coefficients in the mixtures and the non-linear parameters of the features. The optimization problems are composed of a data fidelity term and a (ℓ_1, L^p)-penalty. We prove high probability bounds on the prediction errors associated to our estimators. The proof is based on the existence of certificate functions. Following recent works on the geometry of off-the-grid methods, we show that such functions can be constructed provided the parameters of the active features are pairwise separated by a constant with respect to a Riemannian metric. When the number of signals is finite and the noise is assumed Gaussian, we give refinements of our results for $p = 1$ and $p = 2$ using tail bounds on suprema of Gaussian and χ^2 random processes. When $p = 2$, our prediction error reaches the rates obtained by the Group-Lasso estimator in the multi-task linear regression model.

The material of this chapter has been released in [Butucea et al., 2022c].

3.1 Introduction

Observing repeatedly the same process is very frequent nowadays, due to the abundance of data in all fields. Multi-task learning considers the simultaneous analysis of multiple datasets and produces an estimator for each dataset. Datasets can be either discrete-time (e.g. regression models) or continuous-time in our context. We assume that they bring information on the same underlying structure, but can also be contaminated at some extent by outliers.

We assume each process has a signal-plus-noise structure and that the signal is a mixture of features issued from a dictionary of smooth functions parametrized by some non-linear parameter (such as location, scale, etc.). Such mixtures can be seen e.g. in spectroscopy where each feature corresponds to a chemical component of the analyzed material, see [Butucea et al., 2021].

We are interested in recovering simultaneously the signals, i.e. the linear weights in the mixture and the non-linear parameters of the features, by minimizing a weighted prediction risk penalized by the sum of the total energy of the weights that each feature has through the collection of all processes. The prediction risk may put more weight on prescribed signals of interest. We give high probability bounds on the weighted prediction risk that are analogous to the case of multi-task discrete linear regression models.

3.1.1 Model and method

Let $(\mathcal{Z}, \mathcal{F}, \nu)$ be a measure space with ν a finite positive non-zero measure and let H_T be a Hilbert space where the parameter $T \in \mathbb{N}$ accounts for the increasing asymptotic information in the model. The Hilbert space H_T is endowed with the scalar product $\langle \cdot, \cdot \rangle_T$ and the norm $\|\cdot\|_T$. We shall consider the space $L_T = L^2(\nu, H_T)$, the set of H_T-valued strong measurable functions f defined on $(\mathcal{Z}, \mathcal{F}, \nu)$ such that $\|f\|_{L_T} = \sqrt{\int_{\mathcal{Z}} \|f(z)\|_T^2 \, \nu(\mathrm{d}z)}$ is finite. We then endow L_T with a scalar product noted $\langle \cdot, \cdot \rangle_{L_T}$ defined for any $f, g \in L_T$ by :

$$\langle f, g \rangle_{L_T} = \int_{\mathcal{Z}} \langle f(z), g(z) \rangle_T \, \nu(\mathrm{d}z).$$

The norm $\|\cdot\|_{L_T}$ is the natural norm associated with the scalar product and L_T is a Hilbert space, see [Diestel and Uhl, 1977, Section IV]. For $p \in [1, +\infty)$, we write $L^p(\nu, \mathbb{R}^K)$ for the space of $\mathbb{R}^K$-valued measurable function f defined on $(\mathcal{Z}, \mathcal{F}, \nu)$ such that

$$\|f\|_{L^p(\nu, \mathbb{R}^K)} = \left(\int_{\mathcal{Z}} \|f(z)\|_{\ell_2}^p \, \nu(\mathrm{d}z) \right)^{\frac{1}{p}}$$

is finite, where $\|\cdot\|_{\ell_2}$ is the usual Euclidean norm on $\mathbb{R}^K$. We simply write $L^p(\nu)$ for $L^p(\nu, \mathbb{R})$.

We assume we observe a random element Y of the Hilbert space L_T. For any $z \in \mathcal{Z}$, the element $Y(z) \in H_T$ has a signal-plus-noise structure. The signal part is a mixture (linear combination) of smooth features $\varphi_T(\theta)$ belonging to H_T and continuously parametrized by a real parameter $\theta \in \Theta \subseteq \mathbb{R}$. Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space, we note W_T the additional noise process defined on this space and assumed to be almost surely an element of L_T. We denote by $(\varphi_T(\theta), \theta \in \Theta)$ the continuous dictionary formed by all the features. For all $z \in \mathcal{Z}$, we note $\mathcal{Q}^\star(z)$ the finite set of the parameters of the active features appearing in $Y(z)$. We assume that the unknown number of active features s in the observation Y is bounded by a constant K, that is:

$$K \geq \mathrm{Card}\Big(\bigcup_{z \in \mathcal{Z}} \mathcal{Q}^\star(z) \Big) := s. \tag{3.1}$$

In the following we make a slight abuse of notation by writing $\mathcal{Q}^\star$ instead of $\bigcup_{z \in \mathcal{Z}} \mathcal{Q}^\star(z)$.

We consider features $\varphi_T(\theta)$ that are non degenerate, *i.e.* for any $\theta \in \Theta$, $\|\varphi_T(\theta)\|_T$ is finite and non-zero. Let us define the normalized function $\phi_T(\theta)$ for $\theta \in \Theta$ and its multivariate counterpart $\Phi_T(\vartheta)$ for $\vartheta = (\theta_1, \cdots, \theta_K) \in \Theta^K$ by :

$$\phi_T(\theta) = \frac{\varphi_T(\theta)}{\|\varphi_T(\theta)\|_T} \quad \text{and} \quad \Phi_T(\vartheta) = \begin{pmatrix} \phi_T(\theta_1) \\ \vdots \\ \phi_T(\theta_K) \end{pmatrix}.$$

We consider the model with unknown parameters $B^\star$ in $L^2(\nu, \mathbb{R}^K)$ and $\vartheta^\star$ in Θ^K:

$$Y = B^\star \Phi_T(\vartheta^\star) + W_T \quad \text{in } L_T. \tag{3.2}$$

In this work, we assume that the application $B^\star : \mathcal{Z} \to \mathbb{R}^K$ is $s-$ sparse that is,

$$1 \leq s < K \text{ with } s = \mathrm{Card}(S^\star) \text{ and } S^\star = \{k,\ \|B^\star_k\|_{L^2(\nu)} \neq 0\ \}.$$

We remark that the model (3.2) is an extension of the model described in [Butucea et al., 2022a], as the latter amounts to taking ν as a Dirac measure. We gain in generality by letting the measure ν be any finite positive non-zero measure on $\mathcal{Z}$. By doing so, we can consider multiple mixture models.

Example 3.1.1 ($\mathcal{Z} = \{1, \cdots, n\}$). The framework presented above covers a large variety of multiple non-linear regression models. Assume we observe $n \in \mathbb{N}$ random elements of a Hilbert space. Assume that each element is a linear combination of features belonging to a continuous dictionary and is corrupted by a noise process. We encompass this model by indexing the n random elements, setting $\mathcal{Z}$ as the set of indices $\{1, \cdots, n\}$ and the measure ν as the counting measure on this set. The n observations in H are then $(Y(i),\ i = 1, \cdots, n)$.

We might be interested in associating to each observation $Y(i)$ a score indicating, for example, the reliability of the method of acquisition of the observed data. In this context, one can add the information to the model by assigning weights $\nu(i)$ to each process $Y(i)$ and average the prediction risk accordingly.

Example 3.1.2 ($\mathcal{Z}$ is a continuum). By letting $(\mathcal{Z}, \mathcal{F}, \nu)$ be any measure space such that $\nu(\mathcal{Z}) < +\infty$, we can take $\mathcal{Z}$ as a compact interval of $\mathbb{R}$ and ν as the Lebesgue measure on $\mathcal{Z}$. Hence, we generalize the "Function-on-Scalar" models that have many applications including in genomics (see [Barber et al., 2017]) by allowing the design matrix to be parametrized. The "Function-on-Scalar" models refer to regression models where the linear coefficients depend on a time or spatial continuous parameter. Thus, the observation $(Y(z), z \in \mathcal{Z})$ are longitudinal data.

In order to perform signal reconstruction, we are interested in recovering the application $B^\star$ with unknown sparsity s restricted to its support, that is $B^\star_{S^\star}$, and the associated parameters $\vartheta^\star_{S^\star}$ of the nonlinear parametric functions involved in the mixture model.

In order to recover the sparse application $B^\star$ as well as the associated parameters $\vartheta^\star_{S^\star}$ (up to a permutation) we solve a regularized optimization problem with a real tuning parameter $\kappa > 0$ and $p \in [1, 2]$:

$$(\hat{B}, \hat{\vartheta}) \in \underset{B \in L^2(\nu, \mathbb{R}^K), \vartheta \in \Theta_T^K}{\operatorname{argmin}} \quad \frac{1}{2\nu(\mathcal{Z})} \|Y - B\Phi_T(\vartheta)\|^2_{L_T} + \kappa \|B\|_{\ell_1, L^p(\nu)}, \tag{3.3}$$

where for $z \mapsto B(z) = (B_1(z), \ldots, B_K(z))$ in $L^2(\nu, \mathbb{R}^K)$:

$$\|B\|_{\ell_1, L^p(\nu)} = \sum_{k=1}^K \|B_k\|_{L^p(\nu)}.$$

The set Θ_T on which the optimization of the non-linear parameters is performed is required to be a compact interval and the function Φ_T is continuous. When $\mathcal{Z}$ is finite, the existence of at least a solution is therefore guaranteed. When $\mathcal{Z}$ is infinite (and $p \in (1, 2]$), we may use the following result whose proof is given in Section 3.9.1.

Proposition 3.1.3. *Let $p \in (1,2]$. Assume that the function $\theta \mapsto \phi_T(\theta)$ is continuous. Then, the minimization problem (3.3) over $L^2(\mathcal{Z}, \mathbb{R}^K) \times \Theta_T^K$, where Θ_T is a compact interval of $\mathbb{R}$, admits at least one solution.*

In the following, we shall assume that $p \in [1,2]$. This will allow us to control norms of elements in the dual space $L^q(\nu)$ of $L^p(\nu)$, where $1/p + 1/q = 1$, using that $L^q(\nu) \subset L^p(\nu)$ as $p \leq q$.

In this chapter, we aim at quantifying the quality of the prediction of $B^\star \Phi(\vartheta^\star)$ by $\hat{B}\Phi(\hat{\vartheta})$ for $\hat{B}$ and $\hat{\vartheta}$ given by (3.3), by providing an upper bound with high probability of the squared prediction error:

$$\hat{R}_T^2 = \frac{1}{\nu(\mathcal{Z})} \left\| B^\star \Phi(\vartheta^\star) - \hat{B}\Phi(\hat{\vartheta}) \right\|_{L_T}^2. \tag{3.4}$$

Example 3.1.4. Let us set as an example $\mathcal{Z} = \{1, \cdots, n\}$ and ν the counting measure $\sum_{i=1}^n \delta_i$. Assume the n observations belong to the Hilbert space $L^2(\lambda)$ for some measure λ (either discrete or continuous) on the Borel sigma field of $\mathbb{R}$. In this case, the squared prediction error becomes:

$$\hat{R}_T^2 = \frac{1}{n} \sum_{i=1}^n \left\| B^\star(i) \Phi(\vartheta^\star) - \hat{B}(i)\Phi(\hat{\vartheta}) \right\|_{L^2(\lambda)}^2.$$

3.1.2 Previous work

Reconstructing from observations (that are discrete or continuous-time processes) signals that are linear combinations of features belonging to a continuous dictionary $(\varphi(\theta), \theta \in \Theta)$ has applications in many fields such as spectrocopy ([Butucea et al., 2021]), microscopy ([Denoyelle et al., 2020]), super-resolution ([Candès and Fernandez-Granda, 2014]) or spike deconvolution ([Duval and Peyré, 2015]).

Most often, the Hilbert space H_T, to which the observations belong, is assumed to be of finite dimension and the dictionary of features is assumed finite of size K. Over the past two decades, the problem of retrieving a sparse vector in the framework of high dimensional regression models ($K \gg \dim(H_T)$) has generated a large number of works ([Tibshirani, 1996], [Bickel et al., 2009], [Bunea et al., 2007], [Candès and Tao, 2007], [Bühlmann and van de Geer, 2011] and references therein). The celebrated Lasso estimator, popularized by [Tibshirani, 1996] and defined by an optimization problem composed of a data fidelity term and a ℓ_1 penalty, has been extensively studied and has proven to be efficient. In addition, its convex formulation makes its resolution easy to handle (see [Beck and Teboulle, 2009] for a resolution via fast iterative shrinkage-thresholding algorithms). Prediction error bounds and estimation bounds with respect to the ℓ_2 norm have been established for the Lasso under coherence assumptions on the finite dictionary. We refer to [van de Geer and Bühlmann, 2009] for an overview of the coherence assumptions. It turns out that these rates have been proven minimax optimal in [Raskutti et al., 2011]. This means that one cannot find any estimator that achieves faster rates in expected value.

The prediction error bounds obtained for sparse high-dimensional linear models encompass the finite dictionary setting. We consider in this chapter continuous dictionaries. As a consequence, the problem of recontruction is highly non-linear. It might be tempting to address this issue by discretizing the parameter space Θ and getting back to a finite dictionary. However, recent papers have advocated that taking a finite subfamily of a continuous dictionary and using a Lasso estimator to retrieve the linear coefficients of the mixture lead to some issues. In particular, the number of active features in the mixture tends to be overestimated, see [Duval and Peyré, 2017a].

A line of work has emerged around the reconstruction of signals that are mixtures of continuously parametrized features by solving a regularized minimization problem over a space of measures. Indeed, one can readily notice that a mixture of non-linear features $\sum_{k \in S^\star} \beta_k^\star \phi(\theta_k^\star)$ can be written as the application of the linear functional $\mu \mapsto \int \phi(\theta) \mu(\mathrm{d}\theta)$

to the atomic measure $\mu^\star = \sum_{k\in S^\star} \beta_k^\star \delta_{\theta_k^\star}$, where δ_x denotes a Dirac measure located in x. The Beurling Lasso (or BLasso) introduced in [de Castro and Gamboa, 2012] has proven to be efficient to retrieve a sparse measure from its images through linear functionals. We stress that when $\dim(H_T) < +\infty$, there exists a solution to the BLasso made up of at most $\dim(H_T)$ Dirac measures. We refer to [Boyer et al., 2019] and [Duval, 2021] for proofs of this result. For this reason, the BLasso has been used as a counterpart of the classical Lasso for continuous dictionaries. We remark that when H_T is infinite dimensional the BLasso may not have *a priori* an atomic solution. It makes its solutions difficult to interpret in our context. That is why we prefer in this chapter to assume a bound K on the unknown number of features s in order to formulate (3.2). When only one element of H_T is observed (*i.e.* $\mathcal{Z}$ is reduced to a singleton and ν is a Dirac measure), this formulation is equivalent to that of the BLasso restricted to the set of atomic measures of at most K atoms. Efficient numerical methods to solve this problem are available such as modifications of the Frank-Wolfe algorithm ([Denoyelle et al., 2020], [Boyd et al., 2017]) or the Conic Gradient Particle Descent ([Chizat, 2021]). We stress that these methods proceed by seeking a solution that is atomic.

It has been shown that under the assumption of the existence of certificate functions, the BLasso retrieves the exact number of features in a small noise regime ([Candès and Fernandez-Granda, 2014] for a specific dictionary and [Duval and Peyré, 2015] in a more general framework). Regarding prediction error bounds, the research has first focused on mixtures of features issued from a dictionary of complex exponentials parametrized by their frequencies. Much progress has been done in super-resolution using the BLasso with this specific dictionary, see [Candès and Fernandez-Granda, 2014], [Candès and Fernandez-Granda, 2013] in this direction. In [Boyer et al., 2017], the authors showed that the prediction error of the BLasso estimator in this specific case almost reached that of the Lasso estimator provided the frequencies are well separated. They adapted previous results from [Bhaskar et al., 2013] and [Tang et al., 2015] for atomic norm denoising and they extended them to a more general case where the noise level is unknown and needs to be estimated. The authors of the present chapter considered in [Butucea et al., 2022a] the model (3.2) when only one signal is considered ($\mathcal{Z}$ is a singleton and ν is a Dirac measure) and showed that when the one-dimensional parameters of the features are well separated, one can build estimators that lead to a nearly optimal prediction error bound. By nearly optimal, we mean that the prediction error bound obtained in [Butucea et al., 2022a] is of the same order (up to a logarithmic factor) as the minimax bounds obtained in the finite dictionary setting where only linear coefficients are to be retrieved. The result covers a large variety of dictionaries and noises. Let us specify that the separation is expressed with respect to a Riemannian metric following the insightful work of [Poon et al., 2021].

3.1.3 Contributions

We extend the work of [Butucea et al., 2022a] to encompass the case of multiple mixture models. Indeed, we let ν be any finite positive non-zero measure. In the framework of multiple high dimensional linear regressions (ℓ_1, ℓ_p)-norm penalties have been used to retrieve sparsity patterns among the signals. These penalties influence globally the estimations of the signals $(B(i)\Phi(\vartheta^\star), i \in \mathcal{Z})$. Let us mention the (ℓ_1, ℓ_2) mixed norm, used to define the Group-Lasso estimator introduced in [Yuan and Lin, 2006] and that has received significant attention since then (see, [Nardi and Rinaldo, 2008], [Bach, 2008], [Chesneau and Hebiri, 2008],[Huang and Zhang, 2010]). It was shown in [Lounici et al., 2011] that the reconstruction of signals via the Group-Lasso estimator outperfom the reconstruction using the Lasso estimator when the signals share some sparsity pattern. Let us mention the work of [Liu and Zhang, 2008] that provides consistency results and prediction error convergence rates for the general case (ℓ_1, ℓ_p) with $p \in [1, +\infty]$. Estimators obtained from regularized problems via mixed norms have been studied in the context of high dimensional multiple linear regression models but

little has been done for the non-linear extension considered in (3.2). It is therefore natural to find counterpart estimators for the setting of continuous dictionaries. Let us highlight the work of [Golbabaee and Poon, 2022] in which an extension of the BLasso has been proposed in order to address multiple mixture models. The authors extended the result of [Duval and Peyré, 2015] to show exact support recovery results in the small noise regime. They used a penalty that is a convex combination of mixed norms on measures. We remark that when applied to atomic measures these norms reduce to the (ℓ_1, ℓ_1) and (ℓ_1, ℓ_2) norms on the weights of the Dirac measures.

In this chapter, we prove a high-probability upper bound on the prediction error for estimators issued from an optimization problem regularized by a mixed norm $(\ell_1, L^p(\nu))$ with $p \in [1, 2]$ for a wide variety of dictionaries in the general framework where ν can be any finite positive measure. We give refinements of this result when the noise is assumed Gaussian and when the measure ν is discrete. These refined bounds on the prediction error use tail bounds on suprema of Gaussian and χ^2 processes. Our results rely on the existence of certificate functions, see Section 3.4. We also give sufficient conditions for their construction.

3.1.4 Organization of the chapter

In Section 3.2, we formulate assumptions on the model and set some definitions. Section 3.3 presents the main results of this chapter. We start by giving a high probability upper bound on the prediction error in the general case where the measure ν can be any finite measure. Then, we give refinements of this result when the measure ν is a finite weighted sum of Dirac measures and the noise process is assumed Gaussian. In Section 3.4, we present the assumptions on certificate functions that are used to state the high probability upper bound on the prediction error in Section 3.4.1. We give in Section 3.4.2 sufficient conditions to construct such functions. Section 3.5 is dedicated to the proof of the high probability upper bound on the prediction error in the most general framework and Sections 3.6-3.7 give proofs for refinements of this result when ν is a finite sum of weighted Dirac measures and the noise is Gaussian. Section 3.8 is dedicated to the proofs of the results stated in Section 3.4.2 on the existence of certificate functions.

3.1.5 Notation

We shall use for convenience the notation $\lesssim$ and write for two real quantities a and b, $a \lesssim b$ if there exists a positive finite constant C independent of the parameters s, K, T and the measure ν such that $a \leq C\, b$.

We also write for two quantities a, b that $a \asymp b$ if $a \lesssim b$ and $b \lesssim a$.

3.2 Assumptions on the model

In this section, we briefly set some definitions and assumptions that are presented and discussed in more detail in [Butucea et al., 2022a, Sections 3, 4, 5].

3.2.1 Regularity and non-degeneracy assumptions on the features

Let be a fixed parameter $T \in \mathbb{N}$. The features $(\varphi_T(\theta), \theta \in \Theta)$ that form a continuous dictionary are elements of the Hilbert space $(H_T, \langle \cdot, \cdot \rangle_T)$. We shall integrate and differentiate those features with respect to their one-dimensional parameter belonging to the interval Θ of $\mathbb{R}$. To do so, we shall use the notions of Bochner integral and Fréchet derivative. We refer to [Butucea et al., 2022a, Section 3.1] for a short presentation of these objects. We recall that for any function $f : \Theta \mapsto H_T$ differentiable at $\theta \in \Theta$, we have for all $g \in H_T$ that:

$$\partial_\theta \langle f(\theta), g \rangle_T = \langle \partial_\theta f(\theta), g \rangle_T \,.$$

In addition, if f is Bochner integrable on Θ, then for all $g \in H_T$, we have that:

$$\int_\Theta \langle f(\theta), g\rangle_T \,\mathrm{d}\theta = \langle \int_\Theta f(\theta)\,\mathrm{d}\theta, g\rangle_T.$$

We shall require the features to satisfy the following regularity assumption.

Assumption 3.2.1 (Smoothness of φ_T)**.** *We assume that the function $\varphi_T : \Theta \to H_T$ is of class $\mathcal{C}^3$ and $\|\varphi_T(\theta)\|_T > 0$ on Θ.*

Assume that Assumption 3.2.1 holds. Recall that $\phi_T(\theta) = \varphi_T(\theta)/\|\varphi_T(\theta)\|_T$ for all $\theta \in \Theta$. We define the continuous function:

$$g_T(\theta) = \|\partial_\theta \phi_T(\theta)\|_T^2. \tag{3.5}$$

It will be convenient to assume the non-degeneracy of the function g_T.

Assumption 3.2.2 (Positivity of g_T)**.** *Assumption 3.2.1 holds and we have $g_T > 0$ on Θ.*

One can easily show that features are non-degenerate by checking that for any $\theta \in \Theta$ the elements $\varphi_T(\theta)$ and $\partial_\theta \varphi_T(\theta)$ of H_T are linearly independent, see [Butucea et al., 2022a, Lemma 3.1] in this direction.

3.2.2 The kernel and its Riemannian derivatives

In this section, we introduce a function on Θ^2, called kernel, that will quantify the correlation between two features in the dictionary. We shall derive from this kernel a Riemannian metric on the parameter space Θ following [Poon et al., 2021]. This metric will be in particular invariant to a reparametrization of the parameter space.

3.2.2.1 Kernel space and associated Riemannian metric

We shall set a few bases on the notion of kernel and refer to [Butucea et al., 2022a] for further details.

We call kernel a real-valued function defined on Θ^2. Let $\mathcal{K}$ be a symmetric kernel of class $\mathcal{C}^2$ such that the function $g_\mathcal{K}$ defined on the one-dimensional and connected set Θ by:

$$g_\mathcal{K}(\theta) = \partial^2_{x,y}\mathcal{K}(\theta,\theta) \tag{3.6}$$

is positive and locally bounded, where ∂_x (resp. ∂_y) denotes the usual derivative with respect to the first (resp. second) variable.

We derive from the kernel $\mathcal{K}$ the metric $\mathfrak{d}_\mathcal{K}(\theta,\theta')$ between $\theta, \theta' \in \Theta$ by:

$$\mathfrak{d}_\mathcal{K}(\theta,\theta') = |G_\mathcal{K}(\theta) - G_\mathcal{K}(\theta')|, \tag{3.7}$$

where $G_\mathcal{K}$ is a primitive of $\sqrt{g_\mathcal{K}}$. We refer to [Butucea et al., 2022a, Remark 4.1] for details on the connection with Riemannian metrics.

We shall need to differentiate the kernel $\mathcal{K}$ on the manifold $(\Theta, g_\mathcal{K})$. We shall use the covariant derivatives that generalize the classical directional derivative of vector fields on a manifold. Since we only consider the case of a one-dimensional parameter space, the covariant derivatives reduce to simple expressions.

For a real-valued function F defined on Θ^2, we say that F is of class $\mathcal{C}^{0,0}$ on Θ^2 if it is continuous on Θ^2, and of class $\mathcal{C}^{i,j}$ on Θ^2, with $i, j \in \mathbb{N}$, as soon as: F is of class $\mathcal{C}^{0,0}$, and if $i \geq 1$ then the function $\theta \mapsto F(\theta,\theta')$ is of class $\mathcal{C}^i$ on Θ and its derivative $\partial_x F$ is of class $\mathcal{C}^{i-1,j}$ on Θ^2, and if $j \geq 1$ the function $\theta' \mapsto F(\theta,\theta')$ is of class $\mathcal{C}^j$ on Θ and its derivative $\partial_y F$ is of class $\mathcal{C}^{i,j-1}$ on Θ^2. For a real-valued symmetric function F defined on Θ^2 of class $\mathcal{C}^{i,j}$,

we define the covariant derivatives $D_{i,j;\mathcal{K}}[F]$ of order $(i,j) \in \mathbb{N}^2$ recursively by $D_{0,0;\mathcal{K}}[F] = F$ and for $i,j \in \mathbb{N}$, assuming that $g_{\mathcal{K}}$ is of class $\mathcal{C}^{\max(i,j)}$, and $\theta, \theta' \in \Theta$:

$$D_{i+1,j;\mathcal{K}}[F](\theta,\theta') = g_{\mathcal{K}}(\theta)^{\frac{i}{2}} \partial_\theta \left(\frac{D_{i,j;\mathcal{K}}[F](\theta,\theta')}{g_{\mathcal{K}}(\theta)^{\frac{i}{2}}} \right) \quad \text{and} \quad D_{i,j;\mathcal{K}}[F](\theta,\theta') = D_{j,i;\mathcal{K}}[F](\theta',\theta).$$

In particular, we have $D_{0,0;\mathcal{K}}[F] = F$, $D_{1,0;\mathcal{K}} = \partial_x F$, $D_{0,1;\mathcal{K}} = \partial_y F$ and $D_{1,1;\mathcal{K}} = \partial^2_{xy} F$. We shall also consider the following modification of the covariant derivative, for $i,j \in \mathbb{N}$:

$$\tilde{D}_{i,j;\mathcal{K}}[F](\theta,\theta') = \frac{D_{i,j;\mathcal{K}}[F](\theta,\theta')}{g_{\mathcal{K}}(\theta)^{i/2}\, g_{\mathcal{K}}(\theta')^{j/2}}.$$

We have $\tilde{D}_{1,0;\mathcal{K}} \circ \tilde{D}_{0,1;\mathcal{K}} = \tilde{D}_{0,1;\mathcal{K}} \circ \tilde{D}_{1,0;\mathcal{K}}$ and for $i,j \in \mathbb{N}$, assuming that $g_{\mathcal{K}}$ is of class $\mathcal{C}^{\max(i,j)}$:

$$\tilde{D}_{i,j;\mathcal{K}} = \left(\tilde{D}_{1,0;\mathcal{K}}\right)^i \circ \left(\tilde{D}_{0,1;\mathcal{K}}\right)^j.$$

The definitions of covariant derivatives and their modifications cover the case of 1-dimensional functions defined on Θ. For any smooth function f defined on Θ, we shall note $D_{i;\mathcal{K}}[f]$ (resp. $\tilde{D}_{i;\mathcal{K}}[f]$) for $D_{i,0;\mathcal{K}}[F]$ (resp. $\tilde{D}_{i,0;\mathcal{K}}[F]$) where $F : (\theta,\theta') \mapsto f(\theta)$.

For $i,j \in \mathbb{N}$, if $\mathcal{K}$ is of class $\mathcal{C}^{i\vee 1, j\vee 1}$, then we consider the real-valued function defined on Θ^2 by:

$$\mathcal{K}^{[i,j]} = \tilde{D}_{i,j;\mathcal{K}}[\mathcal{K}].$$

In particular, when $\mathcal{K}$ is of class $\mathcal{C}^2$, we have:

$$\mathcal{K}^{[0,0]} = \mathcal{K} \quad \text{and} \quad \mathcal{K}^{[1,1]}(\theta,\theta) = 1. \tag{3.8}$$

3.2.2.2 The kernel associated to the dictionary of features

Let $T \in \mathbb{N}$ be fixed and assume that Assumption 3.2.2 holds. We associate to the dictionary of features $(\varphi_T(\theta), \theta \in \Theta)$ a kernel $\mathcal{K}_T$ on Θ^2 defined by:

$$\mathcal{K}_T(\theta,\theta') = \langle \phi_T(\theta), \phi_T(\theta') \rangle_T = \frac{\langle \varphi_T(\theta), \varphi_T(\theta') \rangle_T}{\|\varphi_T(\theta)\|_T \|\varphi_T(\theta')\|_T}. \tag{3.9}$$

In the following, for an expression A we will often replace the notation $A_{\mathcal{K}_*}$ by A_* where $*$ is T or ∞.

We remark that under Assumptions 3.2.1 and 3.2.2 the definitions (3.5) and (3.6) are consistent by Lemma [Butucea et al., 2022a, Lemma 4.3]. Furthermore, we have that the kernel $\mathcal{K}_T$ is of class $\mathcal{C}^{3,3}$ on Θ^2 and for $i,j \in \{0,\dots,3\}$ and for any $\theta,\theta' \in \Theta$:

$$\mathcal{K}_T^{[i,j]}(\theta,\theta') = \langle \tilde{D}_{i;T}[\phi_T](\theta), \tilde{D}_{j;T}[\phi_T](\theta') \rangle_T, \tag{3.10}$$

$$\sup_{\Theta^2} |\mathcal{K}_T^{[0,0]}| \leq 1, \quad \mathcal{K}_T^{[0,0]}(\theta,\theta) = 1, \quad \mathcal{K}_T^{[1,0]}(\theta,\theta) = 0, \quad \mathcal{K}_T^{[2,0]}(\theta,\theta) = -1 \quad \text{and} \quad \mathcal{K}_T^{[2,1]}(\theta,\theta) = 0. \tag{3.11}$$

In practice, the kernel $\mathcal{K}_T$ may be difficult to handle. It might be convenient to approximate $\mathcal{K}_T$ by a kernel $\mathcal{K}_\infty$ for which some assumptions will be easier to check, see [Butucea et al., 2022a, Section 8] in this direction. We shall give some properties that an approximating kernel $\mathcal{K}_\infty$ must verify. Then we shall define a quantity measuring the precision of the approximation of $\mathcal{K}_T$ by $\mathcal{K}_\infty$ over some compact set $\Theta_T \subseteq \Theta$.

Let us first define for a kernel $\mathcal{K}$ of class $\mathcal{C}^{3,3}$ the function on Θ:

$$h_{\mathcal{K}}(\theta) = \mathcal{K}^{[3,3]}(\theta,\theta). \tag{3.12}$$

The following assumption gathers the properties that an approximating kernel $\mathcal{K}_\infty$ must sastify.

Assumption 3.2.3 (Properties of the asymptotic kernel $\mathcal{K}_\infty$)**.** *The symmetric kernel $\mathcal{K}_\infty$ defined on Θ^2 is of class $\mathcal{C}^{3,3}$, the function g_∞ defined by (3.6) on Θ is positive and locally bounded (as well as of class $\mathcal{C}^2$), and we have $\mathcal{K}_\infty(\theta,\theta) = -\mathcal{K}_\infty^{[2,0]}(\theta,\theta) = 1$ for $\theta \in \Theta$. The set $\Theta_\infty \subseteq \Theta$ is an interval and we have for all $i,j \in \{0,1,2\}$:*

$$m_g := \inf_{\Theta_\infty} g_\infty > 0, \quad L_3 := \sup_{\Theta_\infty} h_\infty < +\infty, \quad \text{and} \quad L_{i,j} := \sup_{\Theta_\infty^2} |\mathcal{K}_\infty^{[i,j]}| < +\infty. \tag{3.13}$$

We stress that the interval Θ_∞ is possibly unbounded contrary to the set Θ_T which is compact.

Under assumption 3.2.3, we derive from the kernel $\mathcal{K}_\infty$ the Riemannian metric $\mathfrak{d}_\infty$ as in (3.7). One can show that the metrics $\mathfrak{d}_T$ and $\mathfrak{d}_\infty$ are strongly equivalent on the compact set Θ_T^2. Indeed, we have:

$$\frac{1}{\rho_T}\,\mathfrak{d}_\infty \le \mathfrak{d}_T \le \rho_T\,\mathfrak{d}_\infty,$$

where ρ_T is a finite positive constant defined by:

$$\rho_T = \max\left(\sup_{\Theta_T}\sqrt{\frac{g_T}{g_\infty}}, \sup_{\Theta_T}\sqrt{\frac{g_\infty}{g_T}}\right).$$

We then give an assumption on the quality of approximation of $\mathcal{K}_T$ by $\mathcal{K}_\infty$. We set:

$$\mathcal{V}_T = \max(\mathcal{V}_T^{(1)}, \mathcal{V}_T^{(2)}) \quad \text{with} \quad \mathcal{V}_T^{(1)} = \max_{i,j\in\{0,1,2\}} \sup_{\Theta_T^2} |\mathcal{K}_T^{[i,j]} - \mathcal{K}_\infty^{[i,j]}| \quad \text{and} \quad \mathcal{V}_T^{(2)} = \sup_{\Theta_T} |h_T - h_\infty|. \tag{3.14}$$

Assumption 3.2.4 (Quality of the approximation)**.** *Let $T \in \mathbb{N}$ be fixed. Assumptions 3.2.2 and 3.2.3 hold, the interval $\Theta_T \subset \Theta_\infty$ is a compact interval, and we have:*

$$\mathcal{V}_T \le L_{2,2} \wedge L_3.$$

3.3 Main Results

3.3.1 General bound on the prediction error

The main goal of this chapter is to bound the prediction error (3.4) associated to the estimators defined in (3.3). We first give a bound that holds with a controled probability in the general case where the penalty of the optimization problem (3.3) is the norm $\|\cdot\|_{\ell_1, L^p(\nu)}$ with $p \in [1,2]$. The bound will be expressed as a function of the tuning parameter κ, the sparsity s, the mass of the measure ν and the parameter of the penalty p. It will stand on an event whose probability is bounded from below by tails of distributions of random variables defined by taking the supremum over the compact set Θ_T and the norm $\|\cdot\|_{L^q(\nu)}$ of real-valued processes indexed on $\mathcal{Z} \times \Theta_T$ of the form:

$$X(z,\theta) = \langle W_T(z), g(\theta)\rangle_T\,,$$

for some smooth functions $g : \Theta_T \to H_T$ related to the dictionary of features and where q is the conjugate of p in the sense that $1/q + 1/p = 1$.

The assumptions on the regularity of the dictionary, the regularity of the limit kernel and the proximity to the limit kernel are the same as those from [Butucea et al., 2022a, Theorem 2.1]. Regarding the noise, we only require that it belongs almost surely to $L^q(\nu, H_T)$. We highlight that the Theorem below is proven under the existence of certificate functions. Those certificates generalize that of [Butucea et al., 2022a, Theorem 2.1]. (In particular, they reduce to those in [Butucea et al., 2022a] when ν is a Dirac measure.) A construction of certificates

has been proposed in [Golbabaee and Poon, 2022] for the case where ν is the counting measure. Our construction is slightly different and covers the general case where ν can be any finite positive measure, see Remark 3.8.4. We shall give in Section 3.4.2 sufficient conditions for their existence. It turns out that we can construct such certificates provided the elements of the set $\mathcal{Q}^\star$ defined in (3.1) are pairwise separated with respect to a Riemannian metric. We remark that the separation does not depend on the space $(\mathcal{Z}, \mathcal{F}, \nu)$. In particular, in the example where $\mathcal{Z}$ is a finite set of cardinal n, increasing n does not improve or deteriorate the separation.

We state the main result of this chapter that is proved in Section 3.5.

Theorem 3.3.1. *Let $T \in \mathbb{N}$. Let be $p \in [1,2]$ and $q \in [2,+\infty]$ such that $1/p + 1/q = 1$. When $p = 1$, we assume that $\mathcal{Z}$ is finite. Assume we observe the random element Y of L_T under the regression model (3.2) with a noise W_T belonging to $L^q(\nu, H_T)$ almost surely and unknown parameters $B^\star \in L^2(\nu, \mathbb{R}^K)$ and $\vartheta^\star = (\theta_1^\star, \cdots, \theta_K^\star)$ a vector with entries in Θ_T (compact interval of $\mathbb{R}$). Let us suppose that the following assumptions hold :*

(i) ***Regularity of the dictionary*** *φ_T: The dictionary function φ_T satisfies the smoothness conditions 3.2.1 . The function g_T satisfies the positivity condition 3.2.2.*

(ii) ***Regularity of the limit kernel:*** *The kernel $\mathcal{K}_\infty$ and the functions g_∞ and h_∞, defined on an interval $\Theta_\infty \subset \Theta$, satisfy the smoothness conditions of Assumption 3.2.3.*

(iii) ***Proximity to the limit kernel:*** *The kernel $\mathcal{K}_T$ defined from the dictionary is sufficiently close to the limit kernel $\mathcal{K}_\infty$ in the sense that Assumption 3.2.4 holds.*

(iv) ***Existence of certificates:*** *The non-empty set of unknown parameters $\mathcal{Q}^\star = \{\theta_k^\star,\, k \in S^\star\}$, with $S^\star = \{k,\, \|B_k^\star\|_{L^2(\nu)} \neq 0\ \}$, satisfies Assumptions 3.4.1 and 3.4.2 with the same $r > 0$.*

Then, there exist finite positive constants $\mathcal{C}$, $\mathcal{C}_0$ depending on r and on the kernel $\mathcal{K}_\infty$ defined on Θ_∞ such that we have the prediction error bound of the estimators $\hat{B}$ and $\hat{\vartheta}$ defined for a tuning parameter $\kappa > 0$ (in (3.3)) given by:

$$\frac{1}{\sqrt{\nu(\mathcal{Z})}} \left\| \hat{B}\Phi_T(\hat{\vartheta}) - B^\star \Phi_T(\vartheta^\star) \right\|_{L_T} \leq \mathcal{C}_0 \sqrt{s}\, \nu(\mathcal{Z})^{\frac{1}{p}}\, \kappa, \tag{3.15}$$

with probability larger than

$$1 - \sum_{i=0}^{2} \mathbb{P}\left(M_i > \mathcal{C}\, \kappa\, \nu(\mathcal{Z})\right), \tag{3.16}$$

where M_i is defined by:

$$M_i = \sup_{\theta \in \Theta_T} \left\| \left\langle W_T, \phi_T^{[i]}(\theta) \right\rangle_T \right\|_{L^q(\nu)}, \quad \text{for } i = 0, 1, 2. \tag{3.17}$$

Remark 3.3.2 (On the choice of κ). We typically choose κ in (3.15) as small as possible giving a global bound on the prediction risk small, such that the event on which the bound stands occurs with a sufficiently large probability.

Remark 3.3.3 (On the dimension K). The bound K on the sparsity s does not appear neither in the upper bound on the prediction error (3.15) nor in the lower bound on the probability (3.16). Thus, it can be taken arbitrarily large. This was already the case in [Butucea et al., 2022a] where $\mathcal{Z}$ is a singleton and ν is a Dirac measure, see Remark 2.4 therein.

3.3.2 Explicit bounds for Gaussian noise and finite number of signals

It is not straightforward to establish tail bounds for the random variables M_i defined in Theorem 3.3.1. However, if the noise process for fixed z in $\mathcal{Z}$ is centered Gaussian, for the cases $p = q = 2$ and $p = 1$ together with $q = +\infty$, this can be done using Rice formulae (see [Azaïs and Wschebor, 2009] for a complete overview of Rice formulae). We shall then deduce bounds for arbitrary values of conjugate pairs (p, q) in $[1, 2]\times[2, \infty]$ using interpolation inequalities.

We will give an explicit lower bound for the probability (3.16). The lower bound will depend on the parameter T and the number of signals $n = \mathrm{Card}(\mathcal{Z})$ assumed to be finite here. Thus, we will be able to give a convergence rate towards zero for the prediction error with respect to these parameters.

In order to use tail bounds for the random variables M_j from Theorem 3.3.1, we state additional assumptions on the noise W_T. As in [Butucea et al., 2022a], we make the following assumption on the noise process W_T, where the decay rate $\Delta_T > 0$ controls the noise variance decay as the parameter T grows and $\sigma > 0$ is the intrinsic noise level.

Assumption 3.3.1 (Admissible noise)**.** *Let $T \in \mathbb{N}$. Assume that the set $\mathcal{Z}$ is finite. The processes $(W_T(z),\, z \in \mathcal{Z})$ are independent copies of a noise process w_T. The noise process w_T belongs to H_T almost surely and, there exist a noise level $\sigma > 0$ and a decay rate $\Delta_T > 0$ such that for all $f \in H_T$ the random variable $\langle f, w_T\rangle_T$ is a centered Gaussian random variable satisfying:*

$$\mathrm{Var}\left(\langle f, w_T\rangle_T\right) \leq \sigma^2\, \Delta_T\, \|f\|_T^2. \tag{3.18}$$

3.3.2.1 The case $p = 2$ and $\mathcal{Z}$ finite

We state a corollary of Theorem 3.3.1 for the specific case where ν is an atomic measure composed of n atoms and the penalty of the optimization problem (3.3) is a mixed $(\ell_1, L^2(\nu))$ norm. The proof is given in Section 3.6.

We denote by $|\Theta_T|_{\mathfrak{d}_T}$ the diameter of the interval Θ_T with respect to the Riemannian metric $\mathfrak{d}_T$ associated to the kernel $\mathcal{K}_T$ and defined in (3.7).

Corollary 3.3.4. *Let $T \in \mathbb{N}$. We fix $p = q = 2$. We assume that $\mathrm{Card}(\mathcal{Z}) = n < +\infty$ and that the measure ν is $\nu = \sum_{z\in\mathcal{Z}} a_z\delta_z$ where δ_z denotes a Dirac measure located in $z \in \mathcal{Z}$ and $(a_z, z \in \mathcal{Z})$ are non-negative real numbers. Assume we observe the random element Y of L_T under the regression model (3.2) with unknown parameters $B^\star$ in $L^2(\nu, \mathbb{R}^K)$ (which can be identified with $\mathbb{R}^{n\times K}$) and $\vartheta^\star = (\theta_1^\star, \cdots, \theta_K^\star)$ a vector with entries in Θ_T, a compact interval of $\mathbb{R}$, such that Points (i)-(iv) of Theorem 3.3.1 are satisfied and the noise process W_T satisfies Assumption 3.3.1 for a noise level $\sigma > 0$ and a decay rate for the noise variance $\Delta_T > 0$.*

Then, there exist finite positive constants $\mathcal{C}_0$, $\mathcal{C}_1$, $\mathcal{C}_2$, depending on the kernel $\mathcal{K}_\infty$ defined on Θ_∞ and on r such that for any $\tau > 1$ and a tuning parameter:

$$\kappa \geq \mathcal{C}_1\sigma\sqrt{\frac{\|a\|_{\ell_\infty}\Delta_T\, n}{\nu(\mathcal{Z})^2}}\left(1 + \sqrt{1 + \frac{\log(\tau)}{n}}\right),$$

where $\|a\|_{\ell_\infty} = \max_{z\in\mathcal{Z}} |a_z|$, we have the following prediction error bound of the estimators $\hat{B}$ and $\hat{\vartheta}$ defined in (3.3):

$$\frac{1}{\sqrt{\nu(\mathcal{Z})}}\left\|\hat{B}\Phi_T(\hat{\vartheta}) - B^\star\Phi_T(\vartheta^\star)\right\|_{L_T} \leq \mathcal{C}_0\sqrt{s\,\nu(\mathcal{Z})}\,\kappa, \tag{3.19}$$

with probability larger than $1 - \mathcal{C}_2\left(\frac{1}{\tau} + \frac{|\Theta_T|_{\mathfrak{d}_T}F(n)}{\sqrt{\tau}}\right)$ with a sequence $F(n) \asymp \sqrt{n}\,\mathrm{e}^{-n/2}$.

Remark 3.3.5 (Comparison to the Group-Lasso estimator). Assume that the Hilbert space $H_T = \mathbb{R}^T$ is endowed with the Euclidean scalar product and Euclidean norm $\|\cdot\|_{\ell_2}$. Let $\mathcal{Z} = \{1, \cdots, n\}$ and let ν be the counting measure on $\mathcal{Z}$, *i.e.* $\nu = \sum_{k=1}^n \delta_k$. Notice that in this setting $L_T = L^2(\nu, H_T)$ is of finite dimension and can be identified with $\mathbb{R}^{n\times T}$. Assume that the observation $Y \in L_T$ comes from the model (3.2) where for any $i \in \{1, \cdots, n\}$, $W_T(i)$ is a Gaussian vector in $\mathbb{R}^T$ with independent entries of variance σ^2. Assume also that the Gaussian vectors $(W_T(i),\, 1 \leq i \leq n)$ are independent. Thus, Assumption 3.3.1 holds with an equality in (3.18) and

$$\Delta_T = 1.$$

We first consider that the parameters $\vartheta^\star$ are known. In this case, the model becomes the classical high-dimensional multiple linear regression model and the Group-Lasso estimator $\hat{B}_L$ can be used to estimate $B^\star$ under coherence assumptions on the finite dictionary made of the rows of the matrix $\Phi^\star = \Phi_T(\vartheta^\star) \in \mathbb{R}^{K\times T}$ (see [Bickel et al., 2009]). The authors of [Lounici et al., 2011] showed that the prediction error associated to the Group-Lasso estimator satisfies the bound:

$$\frac{1}{nT}\sum_{i=1}^n \|(\hat{B}_L(i) - B^\star(i))\Phi^\star\|_{\ell_2}^2 \lesssim \frac{\sigma^2\, s}{T}\left(1 + \frac{\log(K)}{n}\right),$$

with high probability, larger than $1 - 1/K^\gamma$ for some positive constant $\gamma > 0$. Furthermore, in the case where $B^\star$ is an unknown s-sparse application, $\vartheta^\star$ is known and $\Phi^\star$ verifies a coherence property, then lower bounds of order $\sigma^2\, s(1 + \log(K/s)/n)/T$ in expected value can be established. The non-asymptotic prediction lower bounds for the prediction error given in [Lounici et al., 2011] are for $2s < K$:

$$\inf_{\hat{B}} \sup_{B^\star\, s-\text{sparse}} \mathbb{E}\left[\frac{1}{nT}\sum_{i=1}^n \|(\hat{B}(i) - B^\star(i))\Phi^\star\|_{\ell_2}^2\right] \geq C \cdot \frac{\sigma^2\, s}{T}\left(1 + \frac{\log(K/s)}{n}\right),$$

where the infinimum is taken over all the estimators $\hat{B}$ (measurable functions of the obervation Y taking their values in $L^2(\nu, \mathbb{R}^K)$) and for some constant $C > 0$ free of s, K, n and T.

When the linear coefficients $B^\star$ and the parameters $\vartheta^\star$ are unknown, Corollary 3.3.4 gives an upper bound for the prediction risk which is similar to that of the linear case. Consider the estimators from (3.3) with $p = 2$. Assume that the Riemannian diameter of the set Θ_T is bounded by a constant free of T. By squaring (3.19) and then dividing it by T, we obtain from Corollary 3.3.4 with:

$$\kappa = \mathcal{C}_1\sigma\sqrt{\frac{1}{n}}\left(1 + \sqrt{1 + \frac{\log(\tau)}{n}}\right) \quad \text{and } \tau = T^\gamma \quad \text{for some given } \gamma > 0,$$

that with high probability, larger than $1 - C'/T^\gamma - C''F(n)/T^{\gamma/2}$:

$$\frac{1}{nT}\sum_{i=1}^n \left\|\hat{B}(i)\Phi_T(\hat{\vartheta}) - B^\star(i)\Phi_T(\vartheta^\star)\right\|_{\ell_2}^2 \lesssim \frac{\sigma^2\, s}{T}\left(1 + \frac{\log(T)}{n}\right). \tag{3.20}$$

We identify two regimes depending on the ratio $\log(T)/n$. Indeed, when $\log(T)/n \gg 1$ the bound (3.20) behaves as $\frac{\sigma^2\, s\log(T)}{nT}$ and stands with probability that converges towards 1 at the rate $F(n)/T^{\gamma/2}$. On the contrary, when $\log(T)/n \ll 1$ the bound (3.20) is of order $\frac{\sigma^2\, s}{T}$ and stands with probality that converges towards 1 at the rate $1/T^\gamma$.

3.3.2.2 The case $p = 1$ and $\mathcal{Z}$ finite

We apply Theorem 3.3.1 to the particular case $p = 1$. It turns out that for $q = +\infty$, tail bounds for the random variables M_j with $j = 0, 1, 2$ can be established from Rice formulae for smooth Gaussian processes. The following Corollary is proved in Section 3.7.

Corollary 3.3.6. *Let $T \in \mathbb{N}$. We fix $p = 1, q = +\infty$. We assume that* $\mathrm{Card}(\mathcal{Z}) = n < +\infty$ *and that the measure ν is $\nu = \sum_{z\in\mathcal{Z}} a_z \delta_z$ where δ_z denotes a Dirac measure located in $z \in \mathcal{Z}$ and $(a_z, z \in \mathcal{Z})$ are non-negative real numbers. Assume we observe the random element Y of L_T under the regression model (3.2) with unknown parameters $B^\star$ in $L^2(\nu, \mathbb{R}^K)$ (which can be identified with $\mathbb{R}^{n\times K}$) and $\vartheta^\star = (\theta_1^\star, \cdots, \theta_K^\star)$ a vector with entries in Θ_T, a compact interval of $\mathbb{R}$, such that Points (i)-(iv) of Theorem 3.3.1 are satisfied and the noise process W_T satisfies Assumption 3.3.1 for a noise level $\sigma > 0$ and a decay rate for the noise variance $\Delta_T > 0$.*

Then, there exist finite positive constants $\mathcal{C}_0$, $\mathcal{C}_3$, $\mathcal{C}_4$, depending on the kernel $\mathcal{K}_\infty$ defined on Θ_∞ and on r such that for any $\tau > 1$ and a tuning parameter:

$$\kappa \geq \mathcal{C}_3 \sigma \sqrt{\Delta_T \log(\tau)}/\nu(\mathcal{Z}),$$

we have the following prediction error bound of the estimators $\hat{B}$ and $\hat{\vartheta}$ defined in (3.3):

$$\frac{1}{\sqrt{\nu(\mathcal{Z})}} \left\| \hat{B}\Phi_T(\hat{\vartheta}) - B^\star \Phi_T(\vartheta^\star) \right\|_{L_T} \leq \mathcal{C}_0 \sqrt{s}\, \nu(\mathcal{Z})\, \kappa,$$

with probability larger than $1 - \mathcal{C}_4\, n \left(\frac{|\Theta_T| \mathfrak{d}_T}{\tau\sqrt{\log \tau}} \vee \frac{1}{\tau} \right)$.

Remark 3.3.7. When the measure ν is composed of one atom, that is $n = 1$. This result covers that of [Butucea et al., 2022a, Theorem 2.1].

Remark 3.3.8 (Comparison to other estimators). Let us set $H_T = \mathbb{R}^T$, $\mathcal{Z} = \{1, \cdots, n\}$, ν the counting measure and W_T as in Remark 3.3.5 and assume that the Riemannian diameter of the set Θ_T is bounded by a constant free of T. We recall that in this case $\Delta_T = 1$. By considering the estimators built from the optimization problem (3.3) with $p = 1$ and applying Corollary 3.3.6, we get with:

$$\kappa = \mathcal{C}_3 \sigma \sqrt{\Delta_T \log \tau}/n \quad \text{and } \tau = T^{\gamma/2} \quad \text{for some given} \quad \gamma > 1,$$

that, with probability, larger than $1 - C\, n/T^{\gamma/2}$:

$$\frac{1}{n\,T} \sum_{i=1}^{n} \left\| \hat{B}(i)\Phi_T(\hat{\vartheta}) - B^\star(i)\Phi_T(\vartheta^\star) \right\|_{\ell_2}^2 \lesssim \frac{\sigma^2\, s\, \log(T)}{T}. \tag{3.21}$$

We note that this simultaneous estimation procedure gives the same result as estimating separately n signals as in [Butucea et al., 2022a] under the assumption that each signal has sparsity s. Individual estimation can be better for those signals with smaller sparsity than the global one we use here.

In Remark 3.3.5, we showed that by taking $p = 2$ in the optimization problem (3.3) defining the estimators $\hat{B}$ and $\hat{\vartheta}$, we obtain the bound (3.20) for a well chosen tuning parameter κ. When n and T are sufficiently large, we remark that the bound (3.21) is larger than the bound (3.20) established for the estimators from Corollary 3.3.4 and stands with a smaller probability.

3.3.2.3 Arbitrary value of p in [1,2]

For the cases $p = 2$ and $p = 1$ we established tail bounds for the random variables M_i for $i = 0, 1, 2$ in Corollaries 3.3.4 and 3.3.6, respectively. We recall that these random variables are obtained by taking the supremum over the set Θ_T and the $L^q(\nu)$ norm of real-valued processes indexed on $\mathcal{Z} \times \Theta_T$. For the case $p = 1, q = +\infty$ we used a Rice formula for suprema of smooth Gaussian processes, see [Butucea et al., 2022a, Lemma A.2]. For the case $p = q = 2$ we used a Rice formula for suprema of chi-squared processes; see Lemma 3.9.1. Unfortunately, in the more general case where $p \in [1, 2]$ and $q \in [2, +\infty]$, such formulae seem

out of reach. However, we may use the log-convexity of L^q-norms and use the controls we obtained for the cases $p = 1$ and $p = 2$. Indeed for any $f \in L^\infty(\nu)$ and $q \in [2, +\infty]$, we have the inequality:

$$\|f\|_{L^q} \leq \|f\|_{L^2}^{\frac{2}{q}} \|f\|_{L^\infty}^{\frac{q-2}{q}}.$$

Hence, we readily deduce the following inclusion for any bound $M \geq 0$:

$$\{\|f\|_{L^q} > M\} \subset \{\|f\|_{L^\infty} > M\} \cup \{\|f\|_{L^2} > M\}.$$

3.4 Certificates

We present the certificate functions whose existence is required in Theorem 3.3.1. Such functions were introduced for exact reconstruction of signals, see [Candès and Plan, 2011], [Candès and Fernandez-Granda, 2014], [Duval and Peyré, 2015]. Exact recovery results for the simultaneous reconstruction of signals via the Group-BLasso were proved in [Golbabaee and Poon, 2022] using an extension of the certificates from [Duval and Peyré, 2015]. In [Poon et al., 2021], sufficient conditions for the existence of certificate functions were proved for a wide variety of dictionaries. The authors showed that certificates can be built provided the parameters of the features to be retrieved are well separated with respect to a Riemannian metric. This result requires some assumptions on the kernel associated to the dictionary. In particular, the kernel must be local concave on its diagonal, strictly inferior to 1 outside the diagonal and smooth. Their construction was used in [Butucea et al., 2022a] to establish prediction error bounds under similar assumptions on the dictionary but for a one-dimensional parameter space Θ.

In this chapter, we extend the notion of certificates for our context of multiple reconstructions of signals, following the work of [Golbabaee and Poon, 2022]. Let us emphasize that we use a different contruction than [Golbabaee and Poon, 2022], see Remark 3.8.4.

3.4.1 Assumptions on the certificates

In this section, we introduce the assumptions on the certificates. We will give later in Section 3.4.2 an explicit construction and sufficient conditions for these assumptions to hold.

Let $T \in \mathbb{N}$. We denote the closed ball centered at $\theta \in \Theta_T$ with radius r by:

$$\mathcal{B}_T(\theta, r) = \{\theta' \in \Theta_T,\ \mathfrak{d}_T(\theta, \theta') \leq r\} \subseteq \Theta_T.$$

Let $r > 0$ and let $\mathcal{Q}^\star$ be a subset of Θ_T of cardinal s. We call near region of $\mathcal{Q}^\star$ the union of balls $\bigcup_{\theta^\star \in \mathcal{Q}^\star} \mathcal{B}_T(\theta^\star, r)$ and far region the set Θ_T minus the near region: $\Theta_T \setminus \bigcup_{\theta^\star \in \mathcal{Q}^\star} \mathcal{B}_T(\theta^\star, r)$.

Assumption 3.4.1 (Interpolating certificate)**.** *Let $p, q \in [1, +\infty]$ such that $p \leq q$ and $1/p + 1/q = 1$, let $T \in \mathbb{N}$, $s \in \mathbb{N}^*$, $r > 0$ and $\mathcal{Q}^\star$ be a subset of Θ_T of cardinal s. Suppose Assumptions 3.2.1 and 3.2.2 on the dictionary $(\varphi_T(\theta),\ \theta \in \Theta)$ and Assumption 3.2.3 on $\mathcal{K}_\infty$ hold. Suppose that $\mathfrak{d}_T(\theta, \theta') > 2r$ for all $\theta, \theta' \in \mathcal{Q}^\star \subset \Theta_T$. There exist finite positive constants C_N, C'_N, C_F, C_B with $C_F < 1$, depending on r and $\mathcal{K}_\infty$, such that for any measurable application $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ such that for any $\theta^\star \in \mathcal{Q}^\star$, $\|V(\cdot, \theta^\star)\|_{L^q(\nu)} = 1$, there exists an element $P \in L^q(\nu, H_T)$ satisfying:*

(i) For all $\theta^\star \in \mathcal{Q}^\star$ and $\theta \in \mathcal{B}_T(\theta^\star, r)$, we have $\|\langle \phi_T(\theta), P\rangle_T\|_{L^q(\nu)} \leq 1 - C_N\, \mathfrak{d}_T(\theta^\star, \theta)^2$.

(ii) For all $\theta^\star \in \mathcal{Q}^\star$ and $\theta \in \mathcal{B}_T(\theta^\star, r)$, we have:

$$\|\langle \phi_T(\theta), P\rangle_T - V(\cdot, \theta^\star)\|_{L^q(\nu)} \leq C'_N\, \mathfrak{d}_T(\theta^\star, \theta)^2.$$

(iii) For all θ in Θ_T, $\theta \notin \bigcup_{\theta^\star \in \mathcal{Q}^\star} \mathcal{B}_T(\theta^\star, r)$ (far region), we have:

$$\|\langle \phi_T(\theta), P \rangle_T\|_{L^q(\nu)} \leq 1 - C_F.$$

(iv) We have $\|P\|_{L_T} \leq C_B \sqrt{s}\, \nu(\mathcal{Z})^{\frac{1}{2p}-\frac{1}{2q}}$.

We call "interpolating certificate" the real-valued functions defined on $\mathcal{Z} \times \Theta$ by $(z, \theta) \mapsto \langle \phi_T(\theta), P(z) \rangle_T$ where P is an element of $L^q(\nu, H_T)$ satisfying Points *(i)* – *(iv)* from 3.4.1.

We emphazise the interpolating properties of those certificates by noticing that for any $\theta^\star \in \mathcal{Q}^\star$ we have from Point *(ii)* for ν-almost every $z \in \mathcal{Z}$ that:

$$\langle \phi_T(\theta^\star), P(z) \rangle_T = V(z, \theta^\star).$$

In order to establish prediction error bounds another type of certificate functions having different interpolating properties will be needed, see [Candès and Fernandez-Granda, 2013], [Tang et al., 2015], [Boyer et al., 2017] in this direction.

Assumption 3.4.2 (Interpolating derivative certificate)**.** *Let $p, q \in [1, +\infty]$ such that $p \leq q$ and $1/p + 1/q = 1$, let $T \in \mathbb{N}$, $s \in \mathbb{N}^*$, $r > 0$ and $\mathcal{Q}^\star$ be a subset of Θ_T of cardinal s. Suppose Assumption 3.2.1 and 3.2.2 on the dictionary $(\varphi_T(\theta),\, \theta \in \Theta)$ and Assumption 3.2.3 on $\mathcal{K}_\infty$ hold. Suppose that $\mathfrak{d}_T(\theta, \theta') > 2r$ for all $\theta, \theta' \in \mathcal{Q}^\star \subset \Theta_T$. There exist finite positive constants c_N, c_F, c_B depending on r and $\mathcal{K}_\infty$ such that for any measurable application $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ such that for any $\theta^\star \in \mathcal{Q}^\star$, $\|V(\cdot, \theta^\star)\|_{L^q(\nu)} = 1$, there exists an element $Q \in L^q(\nu, H_T)$ satisfying:*

(i) For all $\theta^\star \in \mathcal{Q}^\star$ and $\theta \in \mathcal{B}_T(\theta^\star, r)$, we have:

$$\|\langle \phi_T(\theta), Q \rangle_T - V(\cdot, \theta^\star)\, \mathrm{sign}(\theta - \theta^\star)\, \mathfrak{d}_T(\theta, \theta^\star)\|_{L^q(\nu)} \leq c_N\, \mathfrak{d}_T(\theta^\star, \theta)^2.$$

(ii) For all θ in Θ_T and $\theta \notin \bigcup_{\theta^\star \in \mathcal{Q}^\star} \mathcal{B}_T(\theta^\star, r)$ (far region), we have $\|\langle \phi_T(\theta), Q \rangle_T\|_{L^q(\nu)} \leq c_F$.

(iii) We have $||Q||_{L_T} \leq c_B \sqrt{s}\, \nu(\mathcal{Z})^{\frac{1}{2p}-\frac{1}{2q}}$.

We call "interpolating derivative certificate" the real-valued functions defined on $\mathcal{Z} \times \Theta$ by $(z, \theta) \mapsto \langle \phi_T(\theta), Q(z) \rangle_T$ where Q is an element of $L^q(\nu, H_T)$ satisfying Points *(i)* – *(iii)* from 3.4.2.

We remark that for any $\theta^\star \in \mathcal{Q}^\star$ we deduce from Point *(i)* for ν-almost every $z \in \mathcal{Z}$:

$$\langle \phi_T(\theta^\star), Q(z) \rangle_T = 0.$$

Let us remark that when ν is a Dirac measure, the norm $\|\cdot\|_{L^q(\nu)}$ reduces to an absolute value and Assumptions 3.4.1 and 3.4.2 correspond to Assumptions 6.1 and 6.2 of [Butucea et al., 2022a].

In the following, we shall often write by a slight abuse of notation $f(\theta)$ for $f(\cdot, \theta)$ when considering a function f from $\mathcal{Z} \times \Theta$ to $\mathbb{R}$.

3.4.2 Construction of the certificates

We give in this section sufficient conditions for Assumptions 3.4.1 and 3.4.2 to hold. These assumptions rely on the existence of real-valued functions defined on $\mathcal{Z} \times \Theta$ called certificates and of the form:

$$(z, \theta) \mapsto \langle \phi_T(\theta), P(z) \rangle_T,$$

where P is an element of $L^q(\nu, H_T)$ satisfying some properties.

We shall follow the construction from [Poon et al., 2021, Theorem 2] for interpolating certificates and generalize the contruction of [Candès and Fernandez-Granda, 2013, Lemma 2.7] for interpolating derivative certificates. In [Candès and Fernandez-Granda, 2013, Lemma 2.7], the authors consider certificates that are trigonometric polynomials whereas we are interested here in a more general framework. Furthermore, we remark that the constructions aforementioned only cover the case where ν is a Dirac measure whereas ν can be any finite positive measure in our framework.

Once built, we will then show that our certificates satisfy the properties required in Assumptions 3.4.1 and 3.4.2. The proofs of the results of this section will closely follow the proofs of [Butucea et al., 2022a, Propositions 7.4 and 7.5] that cover the case where ν is a Dirac measure (*i.e.* only one signal is considered).

Similarly to [Butucea et al., 2022a], we shall consider bounded kernels locally concave on the diagonal. We shall also require the kernels to be strictly less than 1 outside their diagonal. In order to state these properties clearly, we define for $T \in \bar{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$ and $r > 0$:

$$\begin{aligned} \varepsilon_T(r) &= 1 - \sup\left\{|\mathcal{K}_T(\theta,\theta')|; \quad \theta,\theta' \in \Theta_T \text{ such that } \mathfrak{d}_T(\theta',\theta) \geq r\right\}, \\ \nu_T(r) &= -\sup\left\{\mathcal{K}_T^{[0,2]}(\theta,\theta'); \quad \theta,\theta' \in \Theta_T \text{ such that } \mathfrak{d}_T(\theta',\theta) \leq r\right\}. \end{aligned} \tag{3.22}$$

The quantities $\varepsilon_T(r)$ and $\nu_T(r)$ defined from the considered kernel $\mathcal{K}_T$ and the set Θ_T will have to be positive for some $r > 0$. The positivity may be difficult to show when $T \in \mathbb{N}$. In order to show the positivity of $\varepsilon_T(r)$ and $\nu_T(r)$, one can rather show the positivity of $\varepsilon_\infty(r)$ and $\nu_\infty(r)$ derived from an approximating kernel easier to handle and use [Butucea et al., 2022a, Lemma 7.1].

We define the set $\Theta^s_{T,\delta} \subset \Theta^s_T$ of vector of parameters of dimension $s \in \mathbb{N}^*$ and separation $\delta > 0$ as:

$$\Theta^s_{T,\delta} = \left\{(\theta_1, \cdots, \theta_s) \in \Theta^s_T : \ \mathfrak{d}_T(\theta_\ell, \theta_k) > \delta \text{ for all distinct } k,\ell \in \{1,\dots,s\}\right\}. \tag{3.23}$$

Let us define for $i,j = 0,1,2$ (assuming the kernel $\mathcal{K}_T$ is smooth enough) and $\vartheta = (\theta_1,\dots,\theta_s) \in \Theta^s_T$ the $s \times s$ matrix:

$$\mathcal{K}_T^{[i,j]}(\vartheta) = \left(\mathcal{K}_T^{[i,j]}(\theta_k,\theta_\ell)\right)_{1\leq k,\ell\leq s}.$$

Let I be the identity matrix of size $s \times s$.

Using the convention $\inf \emptyset = +\infty$, We define:

$$\delta_T(u,s) = \inf\left\{\delta > 0 : \ A_{T,\ell_\infty}(\vartheta) \leq u, \vartheta \in \Theta^s_{T,\delta}\right\}, \tag{3.24}$$

where:

$$\begin{aligned} A_{T,\ell_\infty}(\vartheta) = \max\Bigg(&\left\|I - \mathcal{K}_T^{[0,0]}(\vartheta)\right\|_{\mathrm{op},\ell_\infty}, \left\|I - \mathcal{K}_T^{[1,1]}(\vartheta)\right\|_{\mathrm{op},\ell_\infty}, \left\|I + \mathcal{K}_T^{[2,0]}(\vartheta)\right\|_{\mathrm{op},\ell_\infty}, \\ &\left\|\mathcal{K}_T^{[1,0]}(\vartheta)\right\|_{\mathrm{op},\ell_\infty}, \left\|\mathcal{K}_T^{[0,1]}(\vartheta)\right\|_{\mathrm{op},\ell_\infty}, \left\|\mathcal{K}_T^{[1,2]}(\vartheta)\right\|_{\mathrm{op},\ell_\infty}\Bigg), \end{aligned} \tag{3.25}$$

and $\|\cdot\|_{\mathrm{op},\ell_\infty}$ denotes the operator norm associated to the sup-norm $\|\cdot\|_{\ell_\infty}$, that is for a matrix $A \in \mathbb{R}^{s\times s}$,

$$\|A\|_{\mathrm{op},\ell_\infty} = \sup_{x\in\mathbb{R}^s, \|x\|_{\ell_\infty}\leq 1} \|Ax\|_{\ell_\infty}.$$

We define quantities which depend on $\mathcal{K}_\infty$, Θ_∞ and on real parameters $r > 0$ and $\rho \geq 1$:

$$\begin{aligned} H^{(1)}_\infty(r,\rho) &= \frac{1}{2} \wedge L_{2,0} \wedge L_{2,1} \wedge \frac{\nu_\infty(\rho r)}{10} \wedge \frac{\varepsilon_\infty(r/\rho)}{10}, \\ H^{(2)}_\infty(r,\rho) &= \frac{1}{6} \wedge \frac{8\,\varepsilon_\infty(r/\rho)}{10(5+2L_{1,0})} \wedge \frac{8\,\nu_\infty(\rho r)}{9(2L_{2,0}+2L_{2,1}+4)}, \end{aligned}$$

where the constants $L_{i,j}$ are defined in (3.13).

We give sufficient conditions for Assumption 3.4.1 to hold. The proof of the following result is given in Section 3.8.1.

Proposition 3.4.1 (Interpolating certificate)**.** *Let $T \in \mathbb{N}$, $s \in \mathbb{N}^*$, $\rho \geq 1$, $r > 0$ and $p, q \in [1, +\infty]$ such that $p \leq q$ and $1/p + 1/q = 1$. We assume that:*

(i) ***Regularity of the dictionary*** *φ_T: Assumptions 3.2.1 and 3.2.2 hold.*

(ii) ***Regularity of the limit kernel*** *$\mathcal{K}_\infty$: Assumption 3.2.3 holds. Furthermore, we have $r \in \left(0, 1/\sqrt{2L_{2,0}}\right)$, and also $\varepsilon_\infty(r/\rho) > 0$ and $\nu_\infty(\rho r) > 0$.*

(iii) **Separation of the non-linear parameters:** *There exists $u_\infty \in \left(0, H_\infty^{(2)}(r, \rho)\right)$ such that:*

$$\delta_\infty(u_\infty, s) < +\infty.$$

(iv) ***Closeness of the metrics*** *$\mathfrak{d}_T$* ***and*** *$\mathfrak{d}_\infty$: We have $\rho_T \leq \rho$.*

(v) ***Proximity of the kernels*** *$\mathcal{K}_T$* ***and*** *$\mathcal{K}_\infty$:*

$$\mathcal{V}_T \leq H_\infty^{(1)}(r, \rho) \quad \text{and} \quad (s-1)\mathcal{V}_T \leq H_\infty^{(2)}(r, \rho) - u_\infty.$$

Then, with the positive constants:

$$C_N = \frac{\nu_\infty(\rho r)}{180}, \quad C'_N = \frac{5}{8}L_{2,0} + \frac{1}{8}L_{2,1} + \frac{1}{2}, \quad C_B = 2 \quad \text{and} \quad C_F = \frac{\varepsilon_\infty(r/\rho)}{10} \leq 1,$$

Assumption 3.4.1 holds (with the same r) for any subset $\mathcal{Q}^\star = \{\theta_i^\star,\, 1 \leq i \leq s\}$ such that for all $\theta \neq \theta' \in \mathcal{Q}^\star$:

$$\mathfrak{d}_T(\theta, \theta') > 2\,\max(r, \rho_T\,\delta_\infty(u_\infty, s)).$$

We state a second result that gives sufficient conditions for Assumption 3.4.2 to hold. The proof is given in Section 3.8.2.

Proposition 3.4.2 (Interpolating derivative certificate)**.** *Let $T \in \mathbb{N}$, $s \in \mathbb{N}^*$ and $p, q \in [1, +\infty]$ such that $p \leq q$ and $1/p + 1/q = 1$. We assume that:*

(i) ***Regularity of the dictionary*** *φ_T: Assumptions 3.2.1 and 2.3.2 hold.*

(ii) ***Regularity of the limit kernel*** *$\mathcal{K}_\infty$: Assumption 3.2.3 holds.*

(iii) ***Separation of the non-linear parameters:*** *There exists $u'_\infty \in (0, 1/6)$, such that:*

$$\delta_\infty(u'_\infty, s) < +\infty.$$

(iv) ***Proximity of the kernels*** *$\mathcal{K}_T$* ***and*** *$\mathcal{K}_\infty$: We have:*

$$\mathcal{V}_T \leq 1 \quad \text{and} \quad (s-1)\mathcal{V}_T + u'_\infty \leq 1/6.$$

Then, with the positive constants:

$$c_N = \frac{1}{8}L_{2,0} + \frac{5}{8}L_{2,1} + \frac{7}{8}, \quad c_B = 2 \quad \text{and} \quad c_F = \frac{5}{4}L_{1,0} + \frac{7}{4},$$

Assumption 3.4.2 holds for any $r > 0$ and any subset $\mathcal{Q}^\star = \{\theta_i^\star,\, 1 \leq i \leq s\}$ such that for all $\theta \neq \theta' \in \mathcal{Q}^\star$:

$$\mathfrak{d}_T(\theta, \theta') > 2\,\max(r, \rho_T\,\delta_\infty(u'_\infty, s)).$$

The assumptions of Proposition 3.4.1 (resp. 3.4.2) are identical to those of [Butucea et al., 2022a, Proposition 7.4] (resp. [Butucea et al., 2022a, Proposition 7.5]). It is not surprising since those results are based on the same construction of certificates. In order to build a certificate $\eta : (z, \theta) \mapsto \mathbb{R}$ satisfying Assumption 3.4.1 or 3.4.2, we shall build for every element $z \in \mathcal{Z}$ certificate functions $\eta_z(\theta) \mapsto \mathbb{R}$ following the same construction as in [Butucea et al., 2022a] and set $\eta(z, \theta) = \eta_z(\theta)$. The functions η_z will be coupled through interpolated values on $\mathcal{Q}^\star$.

3.5 Proof of Theorem 3.3.1

In this section, we shall prove Theorem 3.3.1. We closely follow the proof of [Butucea et al., 2022a, Theorem 2.1] and extend it to the case of a measure ν that is not necessarily a Dirac measure. We decompose the risk over values of estimated non-linear parameters $\hat{\theta}_\ell$ in a neighborhood of the true values $\theta_k^\star$ and those which are far away. Linear functionals of the noise depending on some $\theta \in \Theta_T$ appear in the bounds and we use tail bounds on the suprema of these functionals over all possible values of θ.

Let us bound the prediction error

$$\hat{R}_T := \frac{1}{\sqrt{\nu(\mathcal{Z})}} \left\| \hat{B}\Phi_T(\hat{\vartheta}) - B^\star \Phi_T(\vartheta^\star) \right\|_{L_T}.$$

The predicition error corresponds to the integration on $\mathcal{Z}$ of the prediction error studied in [Butucea et al., 2022a, Theorem 2.1].

By definition (3.3) of $\hat{B}$ and $\hat{\vartheta}$ for the tuning parameter κ, we have:

$$\frac{1}{2\nu(\mathcal{Z})} \left\| Y - \hat{B}\Phi_T(\hat{\vartheta}) \right\|_{L_T}^2 + \kappa \|\hat{B}\|_{\ell_1, L^p(\nu)} \leq \frac{1}{2\nu(\mathcal{Z})} \|Y - B^\star \Phi_T(\vartheta^\star)\|_{L_T}^2 + \kappa \|B^\star\|_{\ell_1, L^p(\nu)}.$$

We define the application $\hat{\Upsilon}$ from L_T to $\mathbb{R}$ by:

$$\hat{\Upsilon}(F) = \left\langle \hat{B}\Phi_T(\hat{\vartheta}) - B^\star \Phi_T(\vartheta^\star), F \right\rangle_{L_T}.$$

This gives, by rearranging terms and using the equation of the model $Y = B^\star \Phi_T(\vartheta^\star) + W_T$, that:

$$\frac{1}{2}\hat{R}_T^2 \leq \frac{1}{\nu(\mathcal{Z})} \hat{\Upsilon}(W_T) + \kappa \left(\|B^\star\|_{\ell_1, L^p(\nu)} - \|\hat{B}\|_{\ell_1, L^p(\nu)} \right). \tag{3.26}$$

Next, we shall expand the two terms on the right-hand side of (3.26). In the rest of the proof, we fix $r > 0$ so that Assumptions 3.4.1 and 3.4.2 are verified for $\mathcal{Q}^\star$. In particular, for all $k \neq k'$ in the support $S^\star = \{k, \|B_k^\star\|_{L^2(\nu)} \neq 0\}$ we have $\mathfrak{d}_T(\theta_k^\star, \theta_{k'}^\star) > 2r$.

We give the definitions of the sets of indices $\hat{S}$, $\tilde{S}_k(r)$ and $\tilde{S}(r)$ for $k \in S^\star$:

- $\hat{S} = \left\{ \ell : \left\| \hat{B}_\ell \right\|_{L^p(\nu)} \neq 0 \right\}$ the support set of $\hat{B}$ given by the optimization problem (3.3);
- $\tilde{S}_k(r) = \left\{ \ell \in \hat{S} : \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star) \leq r \right\}$ the set of indices ℓ in the support of $\hat{B}$ associated to the active parametric functions having $\hat{\theta}_\ell$ close to the true parameter $\theta_k^\star$, for a fixed k in $S^\star$;
- $\tilde{S}(r) = \bigcup_{k \in S^\star} \tilde{S}_k(r)$ the set of indices ℓ in the support of $\hat{B}$ associated to the active parametric functions having $\hat{\theta}_\ell$ close to any true parameter $\theta_k^\star$, for some k in $S^\star$.

Since the closed balls $\mathcal{B}_T(\theta_k^\star, r)$ with $k \in S^\star$ are pairwise disjoint, the sets $\tilde{S}_k(r)$, for $k \in S^\star$, are also pairwise disjoint and one can write the following decomposition with $\tilde{S}(r)^c = \{1, \cdots, K\} \setminus \tilde{S}(r)$:

$$\begin{aligned}
\hat{B}\Phi_T(\hat{\vartheta}) - B^\star \Phi_T(\vartheta^\star) &= \sum_{k=1}^{K} \hat{B}_k \phi_T(\hat{\theta}_k) - \sum_{k \in S^\star} B_k^\star \phi_T(\theta_k^\star) \\
&= \sum_{k \in S^\star, \tilde{S}_k(r) \neq \emptyset} \sum_{\ell \in \tilde{S}_k(r)} \hat{B}_\ell \phi_T(\hat{\theta}_\ell) + \sum_{k \in \tilde{S}(r)^c} \hat{B}_k \phi_T(\hat{\theta}_k) - \sum_{k \in S^\star} B_k^\star \phi_T(\theta_k^\star).
\end{aligned}$$

This decomposition groups the elements of the predicted mixture according to the proximity of the estimated parameter $\hat{\theta}_\ell$ to a true underlying parameter $\theta_k^\star$ to be estimated. We use a Taylor-type expansion with the Riemannian distance $\mathfrak{d}_T$ for the function $\phi_T(\theta)$ around

the elements of $\mathcal{Q}^\star$. By Assumption, the function ϕ_T is twice continuously differentiable with respect to the variable θ and the function g_T is positive on Θ_T. We recall that $\tilde{D}_{i;T}[\phi_T] = \phi_T^{[i]}$ for $i = 0, 1, 2$. According to [Butucea et al., 2022a, Lemma 4.2], we have for any $\theta_k^\star$ and $\hat{\theta}_\ell$ in Θ_T:

$$\phi_T(\hat{\theta}_\ell) = \phi_T(\theta_k^\star) + \operatorname{sign}(\hat{\theta}_\ell - \theta_k^\star)\, \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)\, \phi_T^{[1]}(\theta_k^\star) + \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)^2 \int_0^1 (1-s)\phi_T^{[2]}(\gamma_s^{(k\ell)})\, \mathrm{d}s,$$

where $\gamma^{(k\ell)}$ is a distance realizing geodesic path belonging to Θ_T such that $\gamma_0^{(k\ell)} = \theta_k^\star$, $\gamma_1^{(k\ell)} = \hat{\theta}_\ell$ and $\mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star) = \int_0^1 |\dot{\gamma}_s^{(k\ell)}| \sqrt{g_T(\gamma_s^{(k\ell)})} \mathrm{d}s$.

Hence we obtain:

$$\hat{B}\Phi_T(\hat{\vartheta}) - B^\star \Phi_T(\vartheta^\star) = \sum_{k \in S^\star} I_{0,k}(r)\, \phi_T(\theta_k^\star) + \sum_{k \in S^\star} I_{1,k}(r)\, \phi_T^{[1]}(\theta_k^\star) + \sum_{k \in \tilde{S}(r)^c} \hat{B}_k\, \phi_T(\hat{\theta}_k)$$
$$+ \sum_{k \in S^\star} \left(\sum_{\ell \in \tilde{S}_k(r)} \hat{B}_\ell\, \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)^2 \int_0^1 (1-s)\phi_T^{[2]}(\gamma_s^{(k\ell)})\, \mathrm{d}s \right), \quad (3.27)$$

with

$$I_{0,k}(r) = \left(\sum_{\ell \in \tilde{S}_k(r)} \hat{B}_\ell \right) - B_k^\star \quad \text{and} \quad I_{1,k}(r) = \sum_{\ell \in \tilde{S}_k(r)} \hat{B}_\ell\, \operatorname{sign}(\hat{\theta}_\ell - \theta_k^\star)\, \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star).$$

The functions $I_{0,k}(r)$ and $I_{1,k}(r)$ belong to $L^2(\nu)$. We shall omit the dependence in r when there is no ambiguity. In particular, we write $I_{0,k}(z)$ for $I_{0,k}(r)(z)$. Let us introduce some notations in order to bound the different terms of the expansion above:

$$I_0(r) = \sum_{k \in S^\star} \|I_{0,k}(r)\|_{L^p(\nu)} \qquad \text{and} \qquad I_1(r) = \sum_{k \in S^\star} \|I_{1,k}(r)\|_{L^p(\nu)},$$
$$I_{2,k}(r) = \sum_{\ell \in \tilde{S}_k(r)} \left\|\hat{B}_\ell\right\|_{L^p(\nu)} \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)^2 \qquad \text{and} \qquad I_2(r) = \sum_{k \in S^\star} I_{2,k}(r), \quad (3.28)$$
$$I_3(r) = \sum_{\ell \in \tilde{S}(r)^c} \left\|\hat{B}_\ell\right\|_{L^p(\nu)} = \left\|\hat{B}_{\tilde{S}(r)^c}\right\|_{\ell_1, L^p(\nu)}, \quad (3.29)$$

where $\hat{B}_{\tilde{S}(r)^c}$ denotes the restriction of the vector-valued application $\hat{B}$ to its components in the set of indices $\tilde{S}(r)^c$. We recall that we omit the dependence in r when there is no ambiguity. These quantities are generalizations of the real numbers I_i, where $i = 0, \cdots, 3$, defined in the proof of [Butucea et al., 2022a, Theorem 2.1] as they correspond here to sums of $L^p(\nu)$ norms instead of sums of absolute values.

We bound the difference $\|B^\star\|_{\ell_1, L^p(\nu)} - \left\|\hat{B}\right\|_{\ell_1, L^p(\nu)}$ by noticing that:

$$\|B^\star\|_{\ell_1, L^p(\nu)} - \left\|\hat{B}\right\|_{\ell_1, L^p(\nu)} = \sum_{k \in S^\star} \left(\|B_k^\star\|_{L^p(\nu)} - \sum_{\ell \in \tilde{S}_k(r)} \left\|\hat{B}_\ell\right\|_{L^p(\nu)} \right) - \sum_{k \in \tilde{S}(r)^c} \left\|\hat{B}_k\right\|_{L^p(\nu)} \leq I_0. \quad (3.30)$$

In the next lemma, we give an upper bound of I_0. Recall the constants C_N' and C_F from Assumption 3.4.1.

Let $f \in L^2(\nu)$, we define the application $v : L^2(\nu) \to L^2(\nu)$ such that for any $z \in \mathcal{Z}$:

$$v(f)(z) = \begin{cases} \operatorname{sign}(f(z)) \frac{|f(z)|^{p-1}}{\|f\|_{L^p(\nu)}^{p-1}} & \text{if} \quad \|f\|_{L^p(\nu)} > 0, \\ \nu(\mathcal{Z})^{-\frac{1}{q}} & \text{otherwise,} \end{cases} \quad (3.31)$$

so that $\|v(f)\|_{L^q(\nu)} = 1$.

Lemma 3.5.1. *Under the assumptions of Theorem 3.3.1 and with the element $P_1 \in H_T$ from Assumption 3.4.1 associated to the function $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ defined by:*

$$V(z, \theta_k^\star) = v(I_{0,k})(z),$$

we get that:

$$I_0 \leq C'_N I_2 + (1 - C_F) I_3 + |\hat{\Upsilon}(P_1)|. \tag{3.32}$$

Proof. We have $\|I_{0,k}\|_{L^p(\nu)} = \|I_{0,k}\|^p_{L^p(\nu)} / \|I_{0,k}\|^{p-1}_{L^p(\nu)}$ and therefore:

$$I_0 := \sum_{k \in S^\star} \|I_{0,k}\|_{L^p(\nu)} = \sum_{k \in S^\star} \int V(z, \theta_k^\star) \left(\left(\sum_{\ell \in \tilde{S}_k(r)} \hat{B}_\ell(z) \right) - B_k^\star(z) \right) \nu(\mathrm{d}z).$$

Let P_1 be an element of L_T from Assumption 3.4.1 associated to the function V such that properties $(i)-(iv)$ therein hold. By adding and substracting $\sum_{k \in S^\star} \sum_{\ell \in \tilde{S}_k(r)} \left\langle \hat{B}_\ell \phi_T(\hat{\theta}_\ell), P_1 \right\rangle_{L_T}$ to I_0 and using the property (ii) satisfied by the element P_1, that is, $\langle \phi_T(\theta_k^\star), P_1(z) \rangle_T = V(z, \theta_k^\star)$ for all $k \in S^\star$ and ν-almost every $z \in \mathcal{Z}$, we obtain:

$$I_0 = \sum_{k \in S^\star} \sum_{\ell \in \tilde{S}_k(r)} \int \hat{B}_\ell(z) \left(V(z, \theta_k^\star) - \left\langle \phi_T(\hat{\theta}_\ell), P_1(z) \right\rangle_T \right) \nu(\mathrm{d}z) + \hat{\Upsilon}(P_1) - \sum_{\ell \in \tilde{S}(r)^c} \left\langle \hat{B}_\ell \phi_T(\hat{\theta}_\ell), P_1 \right\rangle_{L_T}.$$

We deduce, using Hölder's inequality, that:

$$I_0 \leq \sum_{k \in S^\star} \sum_{\ell \in \tilde{S}_k(r)} \left\| \hat{B}_\ell \right\|_{L^p(\nu)} \left\| V(\theta_k^\star) - \left\langle \phi_T(\hat{\theta}_\ell), P_1 \right\rangle_T \right\|_{L^q(\nu)} + |\hat{\Upsilon}(P_1)| + \sum_{\ell \in \tilde{S}(r)^c} \left\| \hat{B}_\ell \right\|_{L^p(\nu)} \left\| \left\langle \phi_T(\hat{\theta}_\ell), P_1 \right\rangle_T \right\|_{L^q(\nu)}.$$

Notice that $\hat{\theta}_\ell \notin \bigcup_{k \in S^\star} \mathcal{B}_T(\theta_k^\star, r)$ for $\ell \in \tilde{S}(r)^c$. Then, by using the properties (ii) and (iii) from Assumption 3.4.1, we get that (3.32) holds with the constants C'_N and C_F from Assumption 3.4.1. □

In the next lemma, we give an upper bound of I_1. Recall the constants c_N and c_F from Assumption 3.4.2. Recall the application v defined in (3.31).

Lemma 3.5.2. *Under the assumptions of Theorem 3.3.1 and with the element $Q_0 \in L_T$ from Assumption 3.4.2 associated to the function $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ defined by:*

$$V(z, \theta_k^\star) = v(I_{1,k})(z),$$

we get that:

$$I_1 \leq c_N I_2 + c_F I_3 + |\hat{\Upsilon}(Q_0)|. \tag{3.33}$$

Proof. We have writing $I_{1,k}(z)$ for $I_{1,k}(r)(z)$:

$$I_1 = \sum_{k \in S^\star} \|I_{1,k}\|_{L^p(\nu)} = \sum_{k \in S^\star} \int V(z, \theta_k^\star) I_{1,k}(z) \nu(\mathrm{d}z).$$

Let Q_0 be an element of L_T from Assumption 3.4.2 associated to the function V such that properties $(i)-(iii)$ therein hold. By adding and substracting $\sum_{\ell \in \tilde{S}(r)} \left\langle \hat{B}_\ell \phi_T(\hat{\theta}_\ell), Q_0 \right\rangle_{L_T} =$

$\left\langle \hat{B}\Phi_T(\hat{\vartheta}), Q_0 \right\rangle_{L_T} - \sum_{\ell \in \tilde{S}(r)^c} \left\langle \hat{B}_\ell \phi_T(\hat{\theta}_\ell), Q_0 \right\rangle_{L_T}$ to I_1 and using the triangle inequality, we obtain:

$$I_1 \leq \sum_{k \in S^\star} \sum_{\ell \in \tilde{S}_k(r)} \int |\hat{B}_\ell(z)| \left| V(z, \theta_k^\star)\ \mathrm{sign}(\hat{\theta}_\ell - \theta_k^\star)\, \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star) - \left\langle \phi_T(\hat{\theta}_\ell), Q_0(z) \right\rangle_T \right| \nu(\mathrm{d}z) + \sum_{\ell \in \tilde{S}(r)^c} \left| \left\langle \hat{B}_\ell \phi_T(\hat{\theta}_\ell), Q_0 \right\rangle_{L_T} \right| + \left| \left\langle \hat{B}\Phi_T(\hat{\vartheta}), Q_0 \right\rangle_{L_T} \right|.$$

The property *(i)* of Assumption 3.4.2 gives that $\langle \phi_T(\theta_k^\star), Q_0(z) \rangle_T = 0$ for all $k \in S^\star$ and $\nu-$almost every $z \in \mathcal{Z}$. This implies that $\langle B^\star \Phi_T(\vartheta^\star), Q_0 \rangle_{L_T} = 0$. Then, by using the definition of I_2 and I_3 from (3.28)-(3.29) and the properties *(i)* and *(ii)* of Assumption 3.4.2, we obtain:

$$I_1 \leq c_N I_2 + c_F I_3 + \left| \left\langle \hat{B}\Phi_T(\hat{\vartheta}), Q_0 \right\rangle_{L_T} \right| = c_N I_2 + c_F I_3 + |\hat{\Upsilon}(Q_0)|,$$

with the constants c_N and c_F from Assumption 3.4.2. □

We consider the following random variables for $j = 0, 1, 2$:

$$M_j = \sup_{\theta \in \Theta_T} \left\| \left\langle W_T, \phi_T^{[j]}(\theta) \right\rangle_T \right\|_{L^q(\nu)}.$$

By using the expansion (3.27), Hölder's inequality and the bounds (3.33) and (3.32) for the second inequality, we obtain:

$$\begin{aligned} |\hat{\Upsilon}(W_T)| &\leq (I_0 + I_3) M_0 + I_1 M_1 + I_2\, 2^{-1}\, M_2 \\ &\leq (C_N' I_2 + (2 - C_F) I_3 + |\hat{\Upsilon}(P_1)|) M_0 + (c_N I_2 + c_F I_3 + |\hat{\Upsilon}(Q_0)|) M_1 + I_2\, 2^{-1}\, M_2. \end{aligned} \tag{3.34}$$

At this point, one needs to bound I_2 and I_3. In order to do so, we bound from above and from below the Bregman divergence D_B defined by:

$$D_B = \|\hat{B}\|_{\ell_1, L^p(\nu)} - \|B^\star\|_{\ell_1, L^p(\nu)} - \hat{\Upsilon}(P_0), \tag{3.35}$$

where P_0 is the element given by Assumption 3.4.1 associated to the function V given by:

$$V(z, \theta_k^\star) = \mathrm{sign}(B_k^\star(z)) \frac{|B_k^\star(z)|^{p-1}}{\|B_k^\star\|_{L_p(\nu)}^{p-1}} \quad \text{for all } k \in S^\star. \tag{3.36}$$

The next lemma gives a lower bound of the Bregman divergence.

Lemma 3.5.3. *Under the assumptions of Theorem 3.3.1 and with the constants C_N and C_F of Assumption 3.4.1, we get that:*

$$D_B \geq C_N I_2 + C_F I_3. \tag{3.37}$$

Proof. By definition (3.35) of D_B we have:

$$D_B = \sum_{k \in \hat{S}} \left(\left\| \hat{B}_k \right\|_{L^p(\nu)} - \left\langle \hat{B}_k \phi_T(\hat{\theta}_k), P_0 \right\rangle_{L_T} \right) - \sum_{k \in S^\star} \left(\|B_k^\star\|_{L^p(\nu)} - \langle B_k^\star \phi_T(\theta_k^\star), P_0 \rangle_{L_T} \right).$$

By using the interpolating properties of P_0 from Assumption 3.4.1 associated to V defined in (3.36), we have $\sum_{k\in S^\star} \|B_k^\star\|_{L^p(\nu)} - \langle B_k^\star \phi_T(\theta_k^\star), P_0\rangle_{L_T} = 0$. Hence, we deduce that:

$$\begin{aligned}
D_B &= \sum_{k\in\hat{S}} \left\|\hat{B}_k\right\|_{L^p(\nu)} - \left\langle \hat{B}_k \phi_T(\hat{\theta}_k), P_0 \right\rangle_{L_T} \\
&\geq \sum_{k\in\hat{S}} \left\|\hat{B}_k\right\|_{L^p(\nu)} - \left| \left\langle \hat{B}_k \phi_T(\hat{\theta}_k), P_0 \right\rangle_{L_T} \right| \\
&\geq \sum_{k\in\hat{S}} \left\|\hat{B}_k\right\|_{L^p(\nu)} - \left\|\hat{B}_k\right\|_{L^p(\nu)} \left\| \left\langle \phi_T(\hat{\theta}_k), P_0 \right\rangle_T \right\|_{L^q(\nu)} \\
&\geq \sum_{\ell\in\tilde{S}(r)} \left\|\hat{B}_\ell\right\|_{L^p(\nu)} \left(1 - \left\| \left\langle \phi_T(\hat{\theta}_\ell), P_0 \right\rangle_T \right\|_{L^q(\nu)}\right) \\
&\qquad + \sum_{k\in\tilde{S}(r)^c} \left\|\hat{B}_k\right\|_{L^p(\nu)} \left(1 - \left\| \left\langle \phi_T(\hat{\theta}_k), P_0 \right\rangle_T \right\|_{L^q(\nu)}\right).
\end{aligned}$$

Thanks to properties *(i)* and *(iii)* of Assumption 3.4.1 and the definitions (3.28) and (3.29) of I_2 and I_3, we obtain:

$$D_B \geq \sum_{k\in S^\star} \sum_{\ell\in\tilde{S}_k(r)} C_N \left\|\hat{B}_\ell\right\|_{L^p(\nu)} \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star)^2 + \sum_{k\in\tilde{S}(r)^c} C_F \left\|\hat{B}_k\right\|_{L^p(\nu)} \geq C_N I_2 + C_F I_3,$$

where the constants C_N and C_F are that of Assumption 3.4.1. □

We now give an upper bound of the Bregman divergence.

Lemma 3.5.4. *Under the assumptions of Theorem 3.3.1, we have:*

$$\begin{aligned}
\kappa\,\nu(\mathcal{Z})\, D_B \leq I_2 \left(C_N' M_0 + c_N M_1 + 2^{-1} M_2\right) + I_3 \left((2 - C_F) M_0 + c_F M_1\right) \\
+ |\hat{\Upsilon}(P_1)| M_0 + |\hat{\Upsilon}(Q_0)| M_1 + \kappa\,\nu(\mathcal{Z})\, |\hat{\Upsilon}(P_0)|. \quad (3.38)
\end{aligned}$$

Proof. Recall that $\mathcal{Q}^\star \subset \Theta_T$. We deduce from (3.26) that:

$$\kappa(\|\hat{B}\|_{\ell_1, L^p(\nu)} - \|B^\star\|_{\ell_1, L^p(\nu)}) \leq \frac{1}{\nu(\mathcal{Z})} \hat{\Upsilon}(W_T) - \frac{1}{2}\hat{R}_T^2 \leq \frac{1}{\nu(\mathcal{Z})} \hat{\Upsilon}(W_T).$$

Together with (3.35), we obtain:

$$\kappa D_B \leq \frac{1}{\nu(\mathcal{Z})} |\hat{\Upsilon}(W_T)| + \kappa |\hat{\Upsilon}(P_0)|.$$

Then, use (3.34) to get (3.38). □

By combining the upper and lower bounds (3.37) and (3.38), we deduce that:

$$\begin{aligned}
I_2 \left(C_N - \frac{1}{\kappa\,\nu(\mathcal{Z})} \left(C_N' M_0 + c_N M_1 + 2^{-1} M_2\right)\right) + I_3 \left(C_F - \frac{1}{\kappa\,\nu(\mathcal{Z})} \left((2 - C_F) M_0 + c_F M_1\right)\right) \\
\leq \frac{1}{\kappa\,\nu(\mathcal{Z})} |\hat{\Upsilon}(P_1)| M_0 + \frac{1}{\kappa\,\nu(\mathcal{Z})} |\hat{\Upsilon}(Q_0)| M_1 + |\hat{\Upsilon}(P_0)|. \quad (3.39)
\end{aligned}$$

We define the events:

$$\mathcal{A}_i = \{M_i \leq \mathcal{C}\,\kappa\,\nu(\mathcal{Z})\}, \quad \text{for } i \in \{0, 1, 2\} \quad \text{and} \quad \mathcal{A} = \mathcal{A}_0 \cap \mathcal{A}_1 \cap \mathcal{A}_2, \qquad (3.40)$$

where:

$$\mathcal{C} = \frac{C_F}{2(2 - C_F + c_F)} \wedge \frac{C_N}{2(C'_N + c_N + 2^{-1})}.$$

We get from Inequality (3.39), that on the event $\mathcal{A}$:

$$C_N I_2 + C_F I_3 \leq 2\mathcal{C}' \left(|\hat{\Upsilon}(P_1)| + |\hat{\Upsilon}(Q_0)| + |\hat{\Upsilon}(P_0)| \right) \quad \text{with} \quad \mathcal{C}' = \mathcal{C} \vee 1. \tag{3.41}$$

By reinjecting (3.30), (3.34), (3.32) and (3.33) in (3.26) one gets:

$$\frac{1}{2}\hat{R}_T^2 \leq I_2 \left(\frac{C'_N M_0 + c_N M_1 + 2^{-1} M_2}{\nu(\mathcal{Z})} + \kappa C'_N \right) + I_3 \left(\frac{(2 - C_F) M_0 + c_F M_1}{\nu(\mathcal{Z})} + \kappa(1 - C_F) \right) + |\hat{\Upsilon}(P_1)| \left(\frac{M_0}{\nu(\mathcal{Z})} + \kappa \right) + |\hat{\Upsilon}(Q_0)| \frac{M_1}{\nu(\mathcal{Z})}.$$

Using (3.41), we obtain an upper bound for the prediction error on the event $\mathcal{A}$:

$$\hat{R}_T^2 \leq C\,\kappa\,(|\hat{\Upsilon}(P_0)| + |\hat{\Upsilon}(P_1)| + |\hat{\Upsilon}(Q_0)|), \tag{3.42}$$

with

$$C = 4\mathcal{C}' \left(1 + \frac{\mathcal{C}'}{C_N}(2C'_N + c_N + 1) + \frac{\mathcal{C}'}{C_F}(3 - 2C_F + c_F) \right).$$

Using the Cauchy-Schwarz inequality and the definition of $\hat{\Upsilon}$, we get that for $f \in L_T$:

$$|\hat{\Upsilon}(f)| \leq \hat{R}_T \sqrt{\nu(\mathcal{Z})}\, \|f\|_{L_T}.$$

Using Assumption 3.4.1 *(iv)* for P_0 and P_1, and Assumption 3.4.2 *(iii)* for Q_0, we get:

$$\|P_0\|_{L_T} \leq C_B \sqrt{s} \nu(\mathcal{Z})^{1/2p - 1/2q}, \quad \|P_1\|_{L_T} \leq C_B \sqrt{s} \nu(\mathcal{Z})^{1/2p - 1/2q} \quad \text{and} \quad \|Q_0\|_{L_T} \leq c_B \sqrt{s} \nu(\mathcal{Z})^{1/2p - 1/2q}.$$

Plugging this in (3.42), we get that on the event $\mathcal{A}$:

$$\hat{R}_T^2 \leq \mathcal{C}_0\, \kappa \hat{R}_T\, \sqrt{s}\, \nu(\mathcal{Z})^{\frac{1}{p}} \quad \text{with} \quad \mathcal{C}_0 = (c_B + 2C_B)C.$$

We obtain (3.15) on the event $\mathcal{A}$ defined in (3.40).

3.6 Proof of Corollary 3.3.4

This section is dedicated to the proof of Corollary 3.3.4. We shall apply Theorem 3.3.1 in the particular case $p = 2$ and $q = 2$. Recall that the measure ν is a sum of n weighted Dirac measures. All the assumptions of Theorem 3.3.1 are in force. We shall only give tail bounds for the quantities M_j with $j = 0, 1, 2$ defined in (3.17).

For $j = 0, 1, 2$ and $\theta \in \Theta_T$, we set $X_j(\theta) = \left\| \left\langle W_T, \phi_T^{[j]}(\theta) \right\rangle_T \right\|_{L^2(\nu)}$. Notice that $M_j = \sup_{\Theta_T} X_j$ and that the process X_j^2 is a χ^2 process.

We first consider $j = 0$. Using (3.8) and (3.10), we have that:

$$\|\phi_T(\theta)\|_T^2 = 1 \quad \text{and} \quad \left\|\phi_T^{[1]}(\theta)\right\|_T^2 = \mathcal{K}_T^{[1,1]}(\theta, \theta) = 1.$$

We define two functions f_n and g_n on $\mathbb{R}$ by:

$$f_n(x) = \mathrm{e}^{-x(1 - 2\sqrt{\frac{n}{x}})} \quad \text{and} \quad g_n(x) = \frac{x^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}\, \mathrm{e}^{-x/2}, \tag{3.43}$$

where Γ denotes the gamma function. Notice that both functions are decreasing on $[n, +\infty)$.

We set:

$$A = \frac{\mathcal{C}^2\,\nu(\mathcal{Z})^2}{\sigma^2\|a\|_{\ell_\infty}\Delta_T}.$$

Recall Assumption 3.3.1 on the noise holds. We deduce from Lemma 3.9.2 with $C_1 = C_2 = 1$ and $u = \mathcal{C}^2\,\kappa^2\,\nu(\mathcal{Z})^2$, that for $\kappa \geq \sqrt{(n+1)/A}$:

$$\mathbb{P}\left(M_0^2 > \mathcal{C}^2\,\kappa^2\,\nu(\mathcal{Z})^2\right) \leq f_n\left(\kappa^2\,A\right) + \frac{4|\Theta_T|_{\mathfrak{d}_T}}{2^{n/2}} g_n\left(\kappa^2 A\right), \tag{3.44}$$

where $|\Theta_T|_{\mathfrak{d}_T}$ denotes the diameter of the set Θ_T with respect to the metric $\mathfrak{d}_T$.

We consider $j = 1$. We have by (3.8) and (3.10) that:

$$\left\|\phi_T^{[1]}(\theta)\right\|_T^2 = 1 \quad \text{and} \quad \left\|\tilde{D}_{1;T}[\phi_T^{[1]}](\theta)\right\|_T^2 = \left\|\phi_T^{[2]}(\theta)\right\|_T^2 = \mathcal{K}_T^{[2,2]}(\theta,\theta).$$

Recall $L_{2,2}$ and $\mathcal{V}_T$ are defined in (3.13) and (3.14). Since Assumptions 3.2.3 and 3.2.4 hold, we get that for $\theta \in \Theta_T$:

$$\mathcal{K}_T^{[2,2]}(\theta,\theta) \leq L_{2,2} + \mathcal{V}_T \leq 2L_{2,2}.$$

We deduce from Lemma 3.9.2 with $C_1 = 1$, $C_2 = \sqrt{2L_{2,2}}$ and $u = \mathcal{C}^2\,\kappa^2\,\nu(\mathcal{Z})^2$, that for $\kappa \geq \sqrt{(n+1)/A}$:

$$\mathbb{P}\left(M_1^2 > \mathcal{C}^2\,\kappa^2\,\nu(\mathcal{Z})^2\right) \leq f_n\left(\kappa^2 A\right) + \frac{4\sqrt{2L_{2,2}}\,|\Theta_T|_{\mathfrak{d}_T}}{2^{n/2}} g_n\left(\kappa^2 A\right). \tag{3.45}$$

We consider $j = 2$. We have by (3.10) that:

$$\left\|\phi_T^{[2]}(\theta)\right\|_T^2 = \mathcal{K}_T^{[2,2]}(\theta,\theta) \quad \text{and} \quad \left\|\tilde{D}_{1;T}[\phi_T^{[2]}](\theta)\right\|_T^2 = \left\|\phi_T^{[3]}(\theta)\right\|_T^2 = \mathcal{K}_T^{[3,3]}(\theta,\theta).$$

Recall the definition of the function h_∞ from (3.12) and the constants $L_{2,2}$, L_3, $\mathcal{V}_T$ defined in (3.13) and (3.14). Using also Assumption 3.2.4 so that $\mathcal{V}_T \leq L_{2,2} \wedge L_3$, we get that for all $\theta \in \Theta_T$:

$$\mathcal{K}_T^{[2,2]}(\theta,\theta) \leq L_{2,2} + \mathcal{V}_T \leq 2L_{2,2} \quad \text{and} \quad \mathcal{K}_T^{[3,3]}(\theta,\theta) \leq L_3 + \mathcal{V}_T \leq 2\,L_3.$$

We deduce from Lemma 3.9.2 with $C_1 = \sqrt{2L_{2,2}}$, $C_2 = \sqrt{2L_3}$ and $u = \mathcal{C}^2\,\kappa^2\,\nu(\mathcal{Z})^2$, that for

$$\kappa \geq \sqrt{2\,L_{2,2}\,(n+1)/A},$$

we have:

$$\mathbb{P}\left(M_2^2 > \mathcal{C}^2\,\kappa^2\,\nu(\mathcal{Z})^2\right) \leq f_n\left(\frac{\kappa^2\,A}{2L_{2,2}}\right) + \frac{4\sqrt{L_3}\,|\Theta_T|_{\mathfrak{d}_T}}{\sqrt{L_{2,2}}\,2^{n/2}} g_n\left(\frac{\kappa^2\,A}{2L_{2,2}}\right). \tag{3.46}$$

Wet set:

$$B = \frac{\mathcal{C}_1'^2\,\nu(\mathcal{Z})^2}{\sigma^2\|a\|_{\ell_\infty}\Delta_T} \quad \text{with} \quad \mathcal{C}_1' = \sqrt{\frac{\mathcal{C}^2}{2L_{2,2}\vee 1}}. \tag{3.47}$$

We deduce from (3.44), (3.45) and (3.46) that for $\kappa \geq \sqrt{(n+1)/B}$:

$$\sum_{j=0}^{2} \mathbb{P}(M_j > \mathcal{C}\,\kappa\,\nu(\mathcal{Z})) \leq 3\left(f_n\left(\kappa^2 B\right) + \frac{\mathcal{C}_2'|\Theta_T|_{\mathfrak{d}_T}}{2^{n/2}} g_n\left(\kappa^2 B\right)\right), \tag{3.48}$$

where the constant $\mathcal{C}_2'$ is finite positive and defined by:

$$\mathcal{C}_2' = 4\left(1 \vee \sqrt{2L_{2,2}} \vee \frac{\sqrt{L_3}}{\sqrt{L_{2,2}}}\right).$$

Recall that the functions f_n and g_n are decreasing on $[n, +\infty)$. We get the following asymptotically-equivalent functions (up to a multiplicative constant) for $f_n(c\,n)$ and $g_n(c\,n)$ and some positive constant c:

$$\begin{aligned} f_n(c\,n) &= \mathrm{e}^{-n(c-2\sqrt{c})} \\ g_n(c\,n)/2^{\frac{n}{2}} &\asymp \mathrm{e}^{-\frac{n}{2}(c-\log(c)-1)+\frac{1}{2}\log(n)} \lesssim \mathrm{e}^{-\frac{n}{2}(c-\log(c)-3/2)}\,. \end{aligned} \tag{3.49}$$

Indeed, we use that $\Gamma(n/2) \asymp \mathrm{e}^{\frac{n}{2}\log(\frac{n}{2})-\frac{n}{2}-\frac{1}{2}\log(n)}$. Thus, the constant c determines which of the two terms $f_n(c\,n)$ and $g_n(c\,n)/2^{\frac{n}{2}}$ is dominant.

By solving a second order inequality, we give a lower bound on the tuning parameter κ so that the first right hand term of (3.48) is bounded by $1/\tau$ for some $\tau > 1$.

Indeed for $\tau > 1$ and $\kappa \geq \sqrt{(n+1)/B}\left(1+\sqrt{1+\frac{\log(\tau)}{n}}\right)$ we have:

$$f_n(\kappa^2 B) \leq \frac{1}{\tau}\cdot$$

We also have:

$$g_n(\kappa^2 B) \leq g_n\left(n\left(1+\sqrt{1+\frac{\log(\tau)}{n}}\right)^2\right) \leq g_n\left(n\right)\mathrm{e}^{-n/2}/\sqrt{\tau},$$

where we used that g_n is decreasing on $[n, +\infty)$ for the first inequality and that $\log(1+x) \leq x$ for the second. So that, we get:

$$\sum_{j=0}^{2}\mathbb{P}(M_j > \mathcal{C}\,\kappa\,\nu(\mathcal{Z})) \leq \frac{3}{\tau} + \frac{3\,\mathcal{C}_2'|\Theta_T|_{\mathfrak{d}_T}}{\sqrt{\tau}\,2^{\frac{n}{2}}}g_n\left(n\right)\mathrm{e}^{-n/2}\,.$$

Then, by using (3.49), we deduce an asymptotical equivalence up to a multiplicative constant: $F(n) := g_n(n)\,\mathrm{e}^{-n/2}/2^{n/2} \asymp \mathrm{e}^{-n/2+\log(n)/2}$.

Finally, using the definition of B given in (3.47), when

$$\kappa \geq \mathcal{C}_1\sigma\sqrt{\frac{\|a\|_{\ell_\infty}\Delta_T\,n}{\nu(\mathcal{Z})^2}}\left(1+\sqrt{1+\frac{\log(\tau)}{n}}\right)$$

we get by Theorem 3.3.1 that the bound (3.15) stands with probability larger than $1 - \mathcal{C}_2\left(\frac{1}{\tau}+\frac{|\Theta_T|_{\mathfrak{d}_T}F(n)}{\sqrt{\tau}}\right)$, where:

$$\mathcal{C}_1 = \sqrt{2}/\mathcal{C}_1' \quad \text{and} \quad \mathcal{C}_2 = 3(1 \vee \mathcal{C}_2').$$

This completes the proof of the corollary.

3.7 Proof of Corollary 3.3.6

In this section, we prove Corollary 3.3.6. We shall apply Theorem 3.3.1 in the particular case $p = 1$ and $q = +\infty$. Recall that the measure ν is a sum of n weighted Dirac measures. All the assumptions of Theorem 3.3.1 are in force. We shall only give tail bounds for the quantities M_j with $j = 0, 1, 2$ defined by $M_j = \sup_{\Theta_T} X_j$ where $X_j(\theta) = \left\|\left\langle W_T, \phi_T^{[j]}(\theta)\right\rangle_T\right\|_{L^\infty(\nu)}$.

Using Assumption 3.3.1, we get for any $j = 0, 1, 2$ that:

$$\begin{aligned} \mathbb{P}(M_j > \mathcal{C}\,\kappa\,\nu(\mathcal{Z})) &\leq \sum_{z\in\mathcal{Z}}\mathbb{P}\left(\sup_{\Theta_T}\left\langle W_T(z), \phi_T^{[j]}(\theta)\right\rangle_T > \mathcal{C}\,\kappa\,\nu(\mathcal{Z})\right) \\ &\leq n\mathbb{P}\left(\sup_{\Theta_T}\left\langle w_T, \phi_T^{[j]}(\theta)\right\rangle_T > \mathcal{C}\,\kappa\,\nu(\mathcal{Z})\right). \end{aligned}$$

We use [Butucea et al., 2022a, Lemma A.2] that establishes a tail bound for suprema of smooth Gaussian processes and similar arguments as those developed in the proof of [Butucea et al., 2022a, Theorem 2.1] to get tail bounds on $\sup_{\Theta_T} \left\langle w_T, \phi_T^{[j]}(\theta) \right\rangle_T$ for $j = 0, 1, 2$. We obtain for any $\tau > 1$ and $\kappa \geq \mathcal{C}_3 \sigma \sqrt{\Delta_T \log \tau}/\nu(\mathcal{Z})$ with $\mathcal{C}_3 = \frac{2}{\mathcal{C}} \left(1 \vee \sqrt{2L_{2,2}}\right)$:

$$\mathbb{P}\left(\sup_{\Theta_T} \left\langle w_T, \phi_T^{[j]}(\theta) \right\rangle_T > \mathcal{C}\, \kappa\, \nu(\mathcal{Z}) \right) \leq \mathcal{C}_4' \left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau \sqrt{\log \tau}} \vee \frac{1}{\tau} \right),$$

where $\mathcal{C}_4'$ is a positive constant depending on r and $\mathcal{K}_\infty$ defined in [Butucea et al., 2022a, Eq. (84)]. We get:

$$\sum_{j=0}^{2} \mathbb{P}(M_j > \mathcal{C}\, \kappa\, \nu(\mathcal{Z})) \leq 3\, \mathcal{C}_4'\, n \left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau \sqrt{\log \tau}} \vee \frac{1}{\tau} \right).$$

Therefore, we obtain by Theorem 3.3.1 that (3.15) stands with probability larger than $1 - \mathcal{C}_4\, n \left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau \sqrt{\log \tau}} \vee \frac{1}{\tau} \right)$ with $\mathcal{C}_4 = 3\, \mathcal{C}_4'$ provided the tuning parameter in (3.3) satisfies $\kappa \geq \mathcal{C}_3 \sigma \sqrt{\Delta_T \log \tau}/\nu(\mathcal{Z})$.

3.8 Proofs for the construction of certificates

This section is devoted to the proof of Propositions 3.4.1 and 3.4.2. We shall first introduce norms that will be useful later in the proof. Then, we shall closely follow the proofs of [Butucea et al., 2022a, Propositions 7.4 and 7.5].

Let $p, q \in [1, +\infty]$ such that $p \leq q$ and $1/p + 1/q = 1$, let $m, n \in \mathbb{N}$. We define a norm $\|\cdot\|_{*,q}$ on $L^q(\nu, \mathbb{R}^n)$ by:

$$\|f\|_{*,q} = \max_{1 \leq k \leq n} \|f_k\|_{L^q(\nu)}.$$

We shall also define a norm on any matrix $A \in \mathbb{R}^{n \times m}$ by:

$$\|A\|_{\mathrm{op},*,q} = \sup_{\substack{f \in L^q(\nu, \mathbb{R}^m) \\ \|f\|_{*,q} \leq 1}} \|Af\|_{*,q}.$$

Recall the definition of the operator norm associated to the ℓ_∞ sup-norm defined for any matrix $A \in \mathbb{R}^{n \times m}$ by:

$$\|A\|_{\mathrm{op},\ell_\infty} = \max_{1 \leq k \leq n} \sum_{1 \leq \ell \leq m} |A_{k,\ell}|.$$

We have the following elementary result.

Lemma 3.8.1. *We have the equality on matrix norms on $\mathbb{R}^{n \times m}$:*

$$\|\cdot\|_{\mathrm{op},\ell_\infty} = \|\cdot\|_{\mathrm{op},*,q}.$$

Proof. Let be $A \in \mathbb{R}^{n \times m}$. We have by definition and the triangle inequality for any $f \in L^q(\nu, \mathbb{R}^m)$:

$$\|Af\|_{*,q} = \max_{1 \leq k \leq n} \left\| \sum_{\ell=1}^{m} A_{k,\ell} f_\ell \right\|_{L^q(\nu)} \leq \max_{1 \leq k \leq n} \sum_{\ell=1}^{m} |A_{k,\ell}| \|f_\ell\|_{L^q(\nu)}.$$

Hence for any $f \in L^q(\nu, \mathbb{R}^m)$ such that $\|f\|_{*,q} \leq 1$, we have:

$$\|Af\|_{*,q} \leq \max_{1 \leq k \leq n} \sum_{1 \leq \ell \leq m} |A_{k,\ell}| = \|A\|_{\mathrm{op},\ell_\infty}.$$

Therefore, we have the bound $\|A\|_{\mathrm{op},*,q} \leq \|A\|_{\mathrm{op},\ell_\infty}$.

Let us show that, in fact, we have an equality between those two norms. We set

$$k^\star = \arg\max_{1\le k\le n} \sum_{\ell=1}^{m} |A_{k,\ell}|$$

and we define $f^\star$ so that $f^\star(z) = \nu(\mathcal{Z})^{-1/q}(\mathrm{sign}(A_{k^\star,1}), \cdots, \mathrm{sign}(A_{k^\star,q}))$ for almost every $z \in \mathcal{Z}$. We have $\|f\|_{*,q} = 1$ and $\|Af\|_{*,q} = \|A\|_{\mathrm{op},\ell_\infty}$. Thus, we have $\|A\|_{\mathrm{op},*,q} \ge \|A\|_{\mathrm{op},\ell_\infty}$. Therefore we obtain the equality $\|\cdot\|_{\mathrm{op},\ell_\infty} = \|\cdot\|_{\mathrm{op},*,q}$. □

Since the norm $\|\cdot\|_{\mathrm{op},*,q}$ does not depend on q, we note $\|\cdot\|_{\mathrm{op},*}$ instead of $\|\cdot\|_{\mathrm{op},*,q}$.

Lemma 3.8.2. *Let $x \in \mathbb{R}^m$, $A \in \mathbb{R}^{n\times m}$ and $f \in L^q(\nu, \mathbb{R}^m)$. We have the following inequalities:*

$$\left\|x^\top f\right\|_{L^q(\nu)} \le \|x\|_{\ell_1}\|f\|_{*,q} \quad \text{and} \quad \|Af\|_{*,q} \le \|A\|_{\mathrm{op},*}\|f\|_{*,q}.$$

Proof. This is clear since $\left\|x^\top f\right\|_{L^q(\nu)} \le \sum_{\ell=1}^m |x_\ell|\|f_\ell\|_{L^q(\nu)} \le \|x\|_{\ell_1}\|f\|_{*,q}$.

□

For a function $f : \mathcal{Z} \times \Theta \to \mathbb{R}$, we note for any $z \in \mathcal{Z}$, $f(z)$ (resp. any $\theta \in \Theta$, $f(\theta)$) the function $f(z,\cdot) : \theta \mapsto f(z,\theta)$ (resp. $f(\cdot,\theta) : z \mapsto f(z,\theta)$). The context in which we shall use this notation will be clear so that there is no confusion.

3.8.1 Proof of Proposition 3.4.1(Construction of an interpolating certificate).

Let $T \in \mathbb{N}$ and $s \in \mathbb{N}^*$. Recall Assumptions 3.2.2 (and thus 3.2.1 on the regularity of φ_T) and 3.2.3 on the regularity of the asymptotic kernel $\mathcal{K}_\infty$ are in force. Let $\rho \ge 1$, let $r \in (0, 1/\sqrt{2L_{0,2}})$ and $u_\infty \in \left(0, H_\infty^{(2)}(r,\rho)\right)$ such that *(ii)*, *(iii)*, *(iv)* and *(v)* of Proposition 3.4.1 hold. Recall the definitions (3.23) and (3.24) of $\Theta^s_{T,\delta}$ and δ_∞. By assumption $\delta_\infty(u_\infty, s)$ is finite. Let $\vartheta^\star = (\theta_1^\star, \dots, \theta_s^\star) \in \Theta^s_{T,(2\rho_T\,\delta_\infty(u_\infty,s))\vee(2r)}$. We note $\mathcal{Q}^\star = \{\theta_i^\star,\, 1 \le i \le s\}$ the set of cardinal s. Let $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ such that for any $\theta^\star \in \mathcal{Q}^\star$, $\|V(\theta^\star)\|_{L^q(\nu)} = 1$. Let $\alpha, \xi \in L^q(\nu, \mathbb{R}^s)$. We define the function $P_{\alpha,\xi}$ on $\mathcal{Z}$ as:

$$P_{\alpha,\xi}(z) = \sum_{k=1}^{s} \alpha_k(z)\phi_T(\theta_k^\star) + \sum_{k=1}^{s} \xi_k(z)\, \tilde{D}_{1,T}[\phi_T](\theta_k^\star), \tag{3.50}$$

which belongs to H_T. Recall the definition (3.9) of the kernel $\mathcal{K}_T$. Using (3.10), we define the corresponding certificate function on $\mathcal{Z} \times \Theta$ by:

$$\eta_{\alpha,\xi}(z,\theta) = \langle \phi_T(\theta), P_{\alpha,\xi}(z)\rangle_T = \sum_{k=1}^{s} \alpha_k(z)\, \mathcal{K}_T(\theta,\theta_k^\star) + \sum_{k=1}^{s} \xi_k(z)\, \mathcal{K}_T^{[0,1]}(\theta,\theta_k^\star). \tag{3.51}$$

Notice that the function η is twice continuously differentiable on Θ with respect to its second variable θ due to Assumption 3.2.1. By Assumption 3.2.2 on the regularity of φ_T and the positivity of g_T and (3.10), we get that for almost every $z \in \mathcal{Z}$ the function $\theta \mapsto \eta_{\alpha,\xi}(z,\theta)$ is of class $\mathcal{C}^3$ on Θ, and that:

$$\tilde{D}_{1;T}[\eta_{\alpha,\xi}(z)](\theta) = \sum_{k=1}^{s} \alpha_k(z)\, \mathcal{K}_T^{[1,0]}(\theta,\theta_k^\star) + \sum_{k=1}^{s} \xi_k(z)\, \mathcal{K}_T^{[1,1]}(\theta,\theta_k^\star). \tag{3.52}$$

We give a preliminary technical lemma. Set:

$$\Gamma = \begin{pmatrix} \Gamma^{[0,0]} & \Gamma^{[1,0]\top} \\ \Gamma^{[1,0]} & \Gamma^{[1,1]} \end{pmatrix}, \qquad \text{for } \Gamma^{[i,j]} = \mathcal{K}_T^{[i,j]}(\vartheta^\star). \tag{3.53}$$

As we have $\mathcal{V}(T) \leq \inf_{\Theta_\infty} g_\infty$, by Lemma 7.3 of [Butucea et al., 2022a] we have that:

$$\Theta^s_{T,\rho_T\delta_\infty(u_\infty,s)} \subseteq \Theta^s_{T,\delta_T(u_T(s),s)} \tag{3.54}$$

where $u_T(s) = u_\infty + (s-1)\mathcal{V}_1(T)$. Hence we have:

$$(\theta^\star_i, 1 \leq i \leq s) \in \Theta^s_{T,\delta_T(u_T(s),s)}. \tag{3.55}$$

We deduce from (3.24), (3.25), (3.55) and Lemma 3.8.1 that:

$$\left\|I - \Gamma^{[0,0]}\right\|_{\mathrm{op},*} \leq u_T(s), \quad \left\|I - \Gamma^{[1,1]}\right\|_{\mathrm{op},*} \leq u_T(s), \quad \left\|\Gamma^{[1,0]}\right\|_{\mathrm{op},*} \leq u_T(s)$$
$$\text{and} \quad \left\|\Gamma^{[1,0]\top}\right\|_{\mathrm{op},*} \leq u_T(s). \tag{3.56}$$

We shall write for any $z \in \mathcal{Z}$:

$$\overline{V}(z) = (V(z,\theta^\star_1), \cdots, V(z,\theta^\star_s))^\top. \tag{3.57}$$

Lemma 3.8.3. *Let be $1 \leq p \leq q \leq +\infty$ such that $1/p + 1/q = 1$. Let $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ be a measurable application such that for any $\theta^\star \in \mathcal{Q}^\star$, $\|V(\cdot,\theta^\star)\|_{L^q(\nu)} = 1$. Assume that (3.56) holds. Assume also that $u_T(s) < 1/2$. Then, there exist $\alpha, \xi \in L^q(\nu, \mathbb{R}^s)$ such that:*

$$\eta_{\alpha,\xi}(z,\theta^\star_k) = V(z,\theta^\star_k) \quad \textit{for} \quad 1 \leq k \leq s, \ \textit{for } \nu - \textit{almost every } z, \tag{3.58}$$
$$\tilde{D}_{1,T}[\eta_{\alpha,\xi}(z)](\theta^\star_k) = 0 \quad \textit{for} \quad 1 \leq k \leq s, \ \textit{for } \nu - \textit{almost every } z, \tag{3.59}$$

and we have also that:

$$\|\alpha\|_{*,q} \leq \frac{1-u_T(s)}{1-2u_T(s)}, \quad \|\xi\|_{*,q} \leq \frac{u_T(s)}{1-2u_T(s)}, \quad \left\|\alpha - \overline{V}\right\|_{*,q} \leq \frac{u_T(s)}{1-2u_T(s)},$$

and

$$\|\alpha\|_{*,p} \leq \nu(\mathcal{Z})^{1/p-1/q}\frac{1-u_T(s)}{1-2u_T(s)}, \quad \|\xi\|_{*,p} \leq \nu(\mathcal{Z})^{1/p-1/q}\frac{u_T(s)}{1-2u_T(s)},$$
$$\left\|\alpha - \overline{V}\right\|_{*,p} \leq \nu(\mathcal{Z})^{1/p-1/q}\frac{u_T(s)}{1-2u_T(s)}.$$

Remark 3.8.4. The construction of interpolating certificates is different from the one introduced in [Golbabaee and Poon, 2022] where ν is the counting measure and $q = 2$. Indeed, in [Golbabaee and Poon, 2022] the application ξ is constant and α and ξ solve (3.58) and $\nabla\|\eta_{\alpha,\xi}(\cdot,\theta^\star_k)\|^2_{L^2(\nu)} = 0 \quad \text{for} \quad 1 \leq k \leq s$ instead of (3.59).

Proof. Let $z \in \mathcal{Z}$ such that (3.58) and (3.59) are satisfied. By [Butucea et al., 2022a, Lemma 10.1], we obtain that:

$$\alpha(z) = \Gamma^{-1}_{SC}\overline{V}(z) \quad \text{and} \quad \xi(z) = -[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}\Gamma^{-1}_{SC}\overline{V}(z).$$

where $\Gamma_{SC} = \Gamma^{[0,0]} - \Gamma^{[1,0]\top}[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}$ and:

$$\|I - \Gamma_{SC}\|_{\mathrm{op},*} = \|I - \Gamma_{SC}\|_{\mathrm{op},\ell_\infty} \leq \frac{u_T(s)}{1-u_T(s)}, \quad \left\|\Gamma^{-1}_{SC}\right\|_{\mathrm{op},*} = \left\|\Gamma^{-1}_{SC}\right\|_{\mathrm{op},\ell_\infty} \leq \frac{1-u_T(s)}{1-2u_T(s)}. \tag{3.60}$$

We recall that if M is a matrix such that, $\|I - M\|_{\mathrm{op},*} < 1$, then M is non-singular, $M^{-1} = \sum_{i\geq 0}(I-M)^i$ and $\|M^{-1}\|_{\mathrm{op},*} \leq \left(1 - \|I-M\|_{\mathrm{op},*}\right)^{-1}$. Using (3.56), (3.60), the fact that $\left\|\overline{V}\right\|_{*,q} = 1$ and Lemma 3.8.2, we get:

$$\|\alpha\|_{*,q} \leq \left\|\Gamma_{SC}^{-1}\right\|_{\mathrm{op},*}\left\|\overline{V}\right\|_{*,q} \leq \frac{1-u_T(s)}{1-2u_T(s)},$$

$$\|\xi\|_{*,q} \leq \left\|[\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}\Gamma_{SC}^{-1}\right\|_{\mathrm{op},*}\left\|\overline{V}\right\|_{*,q} \leq \left\|[\Gamma^{[1,1]}]^{-1}\right\|_{\mathrm{op},*}\left\|\Gamma^{[1,0]}\right\|_{\mathrm{op},*}\left\|\Gamma_{SC}^{-1}\right\|_{\mathrm{op},*} \leq \frac{u_T(s)}{1-2u_T(s)},$$

$$\left\|\alpha - \overline{V}\right\|_{*,q} \leq \left\|(\Gamma_{SC}^{-1} - I)\right\|_{\mathrm{op},*}\left\|\overline{V}\right\|_{*,q} \leq \|\Gamma_{SC} - I\|_{\mathrm{op},*}\left\|\Gamma_{SC}^{-1}\right\|_{\mathrm{op},*} \leq \frac{u_T(s)}{1-2u_T(s)}.$$

Then use that for any $f \in L^q(\nu)$, we have

$$\|f\|_{L^p(\nu)} \leq \nu(\mathcal{Z})^{1/p-1/q}\|f\|_{L^q(\nu)} \tag{3.61}$$

by Hölder's inequality as $p \leq q$, to obtain the upper bound on the norm $\|\cdot\|_{*,p}$. This finishes the proof. □

We fix $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ such that for any $\theta^\star \in \mathcal{Q}^\star$ we have $\|V(\theta^\star)\|_{L^q(\nu)} = 1$ and we consider $P_{\alpha,\xi}$ and $\eta_{\alpha,\xi}$ with α and ξ characterized by (3.58) and (3.59) from Lemma 3.8.3. Let $e_\ell \in \mathbb{R}^s$ be the vector with all the entries equal to zero but the ℓ-th which is equal to 1.

Proof of *(iii)* from Assumption 3.4.1 with $C_F = \varepsilon_\infty(r/\rho)/10$. Let $\theta \in \Theta_T$ such that $\mathfrak{d}_T(\theta, \mathcal{Q}^\star) > r$ (far region). It is enough to prove that $\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq 1 - C_F$. Let $\theta_\ell^\star$ be one of the elements of $\mathcal{Q}^\star$ closest to θ in terms of the metric $\mathfrak{d}_T$. Since $\vartheta^\star \in \Theta^s_{T,2\rho_T\delta_\infty(u_\infty,s)}$, we have, by the triangle inequality that for any $k \neq \ell$:

$$2\rho_T\,\delta_\infty(u_\infty, s) < \mathfrak{d}_T(\theta_\ell^\star, \theta_k^\star) \leq \mathfrak{d}_T(\theta_\ell^\star, \theta) + \mathfrak{d}_T(\theta, \theta_k^\star) \leq 2\mathfrak{d}_T(\theta, \theta_k^\star).$$

Hence, we have $\vartheta^\star_{\ell,\theta} \in \Theta^s_{T,\rho_T\delta_\infty(u_\infty,s)}$, where $\vartheta^\star_{\ell,\theta}$ denotes the vector $\vartheta^\star$ whose ℓ-th coordinate has been replaced by θ. Then, we obtain from Lemma 7.3 of [Butucea et al., 2022a] that $\Theta^s_{T,\rho_T\delta_\infty(u_\infty,s)} \subseteq \Theta^s_{T,\delta_T(u_T(s),s)}$ and thus:

$$\vartheta^\star_{\ell,\theta} \in \Theta^s_{T,\delta_T(u_T(s),s)}. \tag{3.62}$$

We denote by $\Gamma_{\ell,\theta}$ (resp. $\Gamma^{[i,j]}_{\ell,\theta}$) the matrix Γ (resp. $\Gamma^{[i,j]}$) in (3.53) where $\vartheta^\star$ has been replaced by $\vartheta^\star_{\ell,\theta}$. Notice the upper bounds (3.56) also hold for $\Gamma_{\ell,\theta}$ because of (3.62). Recall we have that for any $\theta \in \Theta$, $\mathcal{K}_T(\theta,\theta) = 1$ and $\mathcal{K}_T^{[0,1]}(\theta,\theta) = 0$. Elementary calculations give with $\eta_{\alpha,\xi}$ from Lemma 3.8.3 that:

$$\eta_{\alpha,\xi}(z,\theta) = e_\ell^\top\left(\Gamma^{[0,0]}_{\ell,\theta} - I\right)\alpha(z) + \mathcal{K}_T(\theta,\theta_\ell^\star)\alpha_\ell(z) + e_\ell^\top\Gamma^{[1,0]\top}_{\ell,\theta}\xi(z) + \mathcal{K}_T^{[0,1]}(\theta,\theta_\ell^\star)\xi_\ell(z). \tag{3.63}$$

By taking the norm $\|\cdot\|_{L^q(\nu)}$ in (3.63) and using the triangle inequality we get:

$$\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq \left\|\Gamma^{[0,0]}_{\ell,\theta} - I\right\|_{\mathrm{op},*}\|\alpha\|_{*,q} + \|\alpha\|_{*,q}|\mathcal{K}_T(\theta,\theta_\ell^\star)| + \left\|\Gamma^{[1,0]\top}_{\ell,\theta}\right\|_{\mathrm{op},*}\|\xi\|_{*,q} + |\mathcal{K}_T^{[0,1]}(\theta,\theta_\ell^\star)|\|\xi\|_{*,q}. \tag{3.64}$$

Since θ belongs to the "far region", we have by definition of $\varepsilon_T(r)$ given in (3.22) that:

$$|\mathcal{K}_T(\theta,\theta_\ell^\star)| \leq 1 - \varepsilon_T(r).$$

The triangle inequality and the definitions (3.14) of $\mathcal{V}_T$ and (3.13) of $L_{1,0}$ give:

$$|\mathcal{K}_T^{[0,1]}(\theta,\theta_\ell^\star)| \leq L_{1,0} + \mathcal{V}_T. \tag{3.65}$$

Then, using (3.56) (which holds for $\Gamma_{\ell,\theta}$ thanks to (3.62)), we get that:

$$\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq 1 - \varepsilon_T(r) + \frac{u_T(s)}{1-2u_T(s)}\left(2 + L_{1,0} + \mathcal{V}_T\right).$$

Notice that the function $r \mapsto \varepsilon_\infty(r)$ is increasing. Since $\rho_T \leq \rho$, we get by Lemma 7.1 of [Butucea et al., 2022a] that:

$$\varepsilon_T(r) \geq \varepsilon_\infty(r/\rho_T) - \mathcal{V}_T \geq \varepsilon_\infty(r/\rho) - \mathcal{V}_T.$$

By assumption, we have $u_T(s) \leq H^{(2)}_\infty(r,\rho) \leq 1/4$. Hence, we have $\frac{1}{1-2u_T(s)} \leq 2$. We also have $\mathcal{V}_T \leq 1/2$ as $\mathcal{V}_T \leq H^{(1)}_\infty(r,\rho)$. Therefore, we get:

$$\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq 1 - \varepsilon_\infty(r/\rho) + \mathcal{V}_T + u_T(s)\left(5 + 2L_{1,0}\right).$$

The assumption $u_T(s) \leq H^{(2)}_\infty(r,\rho)$ gives:

$$u_T(s) \leq \frac{8}{10\left(5 + 2L_{1,0}\right)}\varepsilon_\infty(r/\rho)\cdot$$

The assumption $\mathcal{V}_T \leq H^{(1)}_\infty(r,\rho)$ gives $\mathcal{V}_T \leq \varepsilon_\infty(r/\rho)/10$. Hence, we have $\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq 1 - \frac{\varepsilon_\infty(r/\rho)}{10}$. Thus, Property *(iii)* from Assumption 3.4.1 holds with $C_F = \varepsilon_\infty(r/\rho)/10$.

Proof of *(i)* from Assumption 3.4.1 with $C_N = \nu_\infty(\rho r)/180$. Let $\theta \in \Theta_T$ such that $\mathfrak{d}_T(\theta, \mathcal{Q}^\star) \leq r$. Let $\ell \in \{1, \cdots, s\}$ such that $\theta \in \mathcal{B}_T(\theta^\star_\ell, r)$ ("near region"). Thus, it is enough to prove that $\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq 1 - C_N\,\mathfrak{d}_T(\theta^\star_\ell, \theta)^2$. This will be done by using Lemma 3.9.3 to obtain a quadratic decay on $\eta_{\alpha,\xi}$ from a bound on its second Riemannian derivative.

Recall that the function $\eta_{\alpha,\xi}$ is twice continuously differentiable with respect to its second variable. Differentiating (3.52) and using that $\mathcal{K}^{[2,0]}_T(\theta,\theta) = -1$ and $\mathcal{K}^{[2,1]}_T(\theta,\theta) = 0$, we deduce that for almost every $z \in \mathcal{Z}$:

$$\tilde{D}_{2;T}[\eta_{\alpha,\xi}(z)](\theta) = e_\ell^\top(I + \Gamma^{[2,0]}_{\ell,\theta})\alpha(z) + \mathcal{K}^{[2,0]}_T(\theta,\theta^\star_\ell)e_\ell^\top\alpha(z) + e_\ell^\top\Gamma^{[2,1]}_{\ell,\theta}\xi(z) + \mathcal{K}^{[2,1]}_T(\theta,\theta^\star_\ell)e_\ell^\top\xi(z). \tag{3.66}$$

We get:

$$\begin{aligned}\tilde{D}_{2;T}[\eta_{\alpha,\xi}(z)](\theta) - V(z,\theta^\star_\ell)\mathcal{K}^{[2,0]}_T(\theta,\theta^\star_\ell) &= e_\ell^\top(I + \Gamma^{[2,0]}_{\ell,\theta})\alpha(z) + \mathcal{K}^{[2,0]}_T(\theta,\theta^\star_\ell)e_\ell^\top(\alpha(z) - \overline{V}(z)) \\ &\quad + e_\ell^\top\Gamma^{[2,1]}_{\ell,\theta}\xi(z) + \mathcal{K}^{[2,1]}_T(\theta,\theta^\star_\ell)e_\ell^\top\xi(z). \end{aligned}\tag{3.67}$$

The triangle inequality and the definition of $\mathcal{V}_T$ give:

$$|\mathcal{K}^{[2,0]}_T(\theta,\theta^\star_\ell)| \leq L_{2,0} + \mathcal{V}_T \quad \text{and} \quad |\mathcal{K}^{[2,1]}_T(\theta,\theta^\star_\ell)| \leq L_{2,1} + \mathcal{V}_T, \tag{3.68}$$

where $L_{2,0}$ and $L_{1,2}$ are defined in (3.13). We deduce from (3.62), the definition of δ_T in (3.24) and (3.25) that:

$$\left\|I + \Gamma^{[2,0]}_{\ell,\theta}\right\|_{\mathrm{op},*} \leq u_T(s) \quad \text{and} \quad \left\|\Gamma^{[2,1]}_{\ell,\theta}\right\|_{\mathrm{op},*} \leq u_T(s). \tag{3.69}$$

We deduce from (3.67) that:

$$\begin{aligned}\left\|\tilde{D}_{2;T}[\eta_{\alpha,\xi}](\theta) - V_\ell(z)\mathcal{K}^{[2,0]}_T(\theta,\theta^\star_\ell)\right\|_{L^q(\nu)} &\leq \|\alpha\|_{*,q}\left\|I + \Gamma^{[2,0]}_{\ell,\theta}\right\|_{\mathrm{op},*} + \left\|\alpha - \overline{V}\right\|_{*,q}(L_{2,0} + \mathcal{V}_T) \\ &\quad + \|\xi\|_{*,q}\left(\left\|\Gamma^{[2,1]}_{\ell,\theta}\right\|_{\mathrm{op},*} + L_{2,1} + \mathcal{V}_T\right) \\ &\leq \frac{u_T(s)}{1-2u_T(s)}(1 + L_{2,0} + L_{2,1} + 2\mathcal{V}_T).\end{aligned}$$

By assumption, we have $u_T(s) \leq H^{(2)}_\infty(r,\rho) \leq 1/6$. Hence, we have $\frac{1}{1-2u_T(s)} \leq 2$. Furthermore, we have by assumption that $\mathcal{V}_T \leq H^{(1)}_\infty(r,\rho) \leq 1/2$ and $u_T(s) \leq H^{(2)}_\infty(r,\rho)$. In particular, we have:

$$u_T(s) \leq \frac{8}{9(2L_{2,0}+2L_{2,1}+4)}\nu_\infty(\rho r).$$

Therefore, we obtain:

$$\left\|\tilde{D}_{2;T}[\eta_{\alpha,\xi}](\theta) - V(z,\theta^\star_\ell)\mathcal{K}^{[2,0]}_T(\theta,\theta^\star_\ell)\right\|_{L^q(\nu)} \leq \frac{8}{9}\nu_\infty(\rho r). \tag{3.70}$$

We now check that the hypotheses of Lemma 3.9.3-*(ii)* hold in order to obtain a quadratic decay on $\theta \mapsto \|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)}$ from the bound (3.70). First recall that for almost every $z \in \mathcal{Z}$, $\theta \mapsto \eta_{\alpha,\xi}(z,\theta)$ is twice continuously differentiable and have the interpolation properties (3.58). By the triangle inequality and since by assumption $\mathcal{V}_T \leq L_{2,0}$, we have:

$$\sup_{\Theta^2_T}|\mathcal{K}^{[2,0]}_T| \leq L_{2,0} + \mathcal{V}_T \leq 2L_{2,0}.$$

Then, Lemma 7.1 of [Butucea et al., 2022a] ensures that for any θ, θ' in Θ_T such that $\mathfrak{d}_T(\theta,\theta') \leq r$ we have:

$$-\mathcal{K}^{[2,0]}_T(\theta,\theta') \geq \nu_\infty(r\rho_T) - \mathcal{V}_T \geq \nu_\infty(\rho r) - \mathcal{V}_T \geq \frac{9}{10}\nu_\infty(\rho r),$$

where we used that the function $r \mapsto \nu_\infty(r)$ is decreasing and $\rho_T \leq \rho$ for the second inequality and that $\mathcal{V}_T \leq H^{(1)}_\infty(r,\rho) \leq \nu_\infty(\rho r)/10$ for the last inequality.

Set $\delta = \frac{8}{9}\nu_\infty(\rho r)$, $\varepsilon = \frac{9}{10}\nu_\infty(\rho r)$, $L = 2L_{2,0}$. As $r < L^{-\frac{1}{2}}$ and $\delta < \varepsilon$, we apply Lemma 3.9.3-*(ii)* and get for $\theta \in \mathcal{B}_T(\theta^\star_\ell, r)$:

$$\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq 1 - \frac{\nu_\infty(\rho r)}{180}\,\mathfrak{d}_T(\theta,\theta^\star_\ell)^2.$$

Proof of *(ii)* from Assumption 3.4.1 with $C'_N = (5L_{2,0}+L_{2,1}+4)/8$. Let $\theta \in \Theta_T$ such that $\mathfrak{d}_T(\theta,\mathcal{Q}^\star) \leq r$. Let $\ell \in \{1,\cdots,s\}$ such that $\theta \in \mathcal{B}_T(\theta^\star_\ell, r)$ ("near region"). We shall prove that $\|\eta_{\alpha,\xi}(\theta) - V(\theta^\star_\ell)\|_{L^q(\nu)} \leq C'_N\,\mathfrak{d}_T(\theta^\star_\ell,\theta)^2$.

Let us consider the function $f : (z,\theta) \to \eta_{\alpha,\xi}(z,\theta) - V(z,\theta^\star_\ell)$. In the following, we will bound $\left\|\tilde{D}_{2;T}[f](\theta)\right\|_{L^q(\nu)}$ on $\mathcal{B}_T(\theta^\star_\ell, r)$ and apply Lemma 3.9.3-*(i)* on f to prove the the inequality of property *(ii)*. Notice that for almost every $z \in \mathcal{Z}$, the map $\theta \mapsto f(z,\theta)$ is twice continuously differentiable. By construction, see (3.58), we have for almost every $z \in \mathcal{Z}$ that $\tilde{D}_{2;T}[f(z)] = \tilde{D}_{2;T}[\eta_{\alpha,\xi}(z)]$, $f(z,\theta^\star_\ell) = 0$ and $\tilde{D}_{1;T}[f(z)](\theta^\star_\ell) = 0$. We deduce from (3.66) and the bounds (3.68) that:

$$\left\|\tilde{D}_{2;T}[f](\theta)\right\|_{L^q(\nu)} \leq \|\alpha\|_{*,q}\left\|I + \Gamma^{[2,0]}_{\ell,\theta}\right\|_{\mathrm{op},*} + \|\alpha\|_{*,q}(L_{2,0}+\mathcal{V}_T) + \|\xi\|_{*,q}\left\|\Gamma^{[2,1]}_{\ell,\theta}\right\|_{\mathrm{op},*} + \|\xi\|_{*,q}(L_{2,1}+\mathcal{V}_T).$$

Using (3.69), and the bounds on α and ξ from Lemma 3.8.3, we get:

$$\left\|\tilde{D}_{2;T}[f](\theta)\right\|_{L^q(\nu)} \leq \frac{1-u_T(s)}{1-2u_T(s)}(L_{2,0}+\mathcal{V}_T+u_T(s)) + \frac{u_T(s)}{1-2u_T(s)}(L_{2,1}+\mathcal{V}_T+u_T(s)).$$

Since $u_T(s) \leq H^{(2)}_\infty(r,\rho) \leq 1/6$ and $\mathcal{V}_T \leq H^{(1)}_\infty(r,\rho) \leq 1/2$, we get:

$$\left\|\tilde{D}_{2;T}[f](\theta)\right\|_{L^q(\nu)} \leq \frac{5}{4}L_{2,0} + \frac{1}{4}L_{2,1} + 1.$$

We get thanks to Lemma 3.9.3-(*i*) on the function f that for any $\theta \in \mathcal{B}_T(\theta^\star_\ell, r)$:

$$\|\eta_{\alpha,\xi}(\theta) - V(\theta^\star_\ell)\|_{L^q(\nu)} \leq \frac{1}{8}\,(5L_{2,0} + L_{1,2} + 4)\;\mathfrak{d}_T(\theta, \theta^\star_\ell)^2.$$

Proof of (*iv*) from Assumption 3.4.1 with $C_B = 2$. Recall the definition of $P_{\alpha,\xi}$ in (3.50). Elementary calculations give using the definitions of $\Gamma^{[0,0]}$ and $\Gamma^{[1,1]}$ in (3.53):

$$\begin{aligned}
\|P_{\alpha,\xi}\|^2_{L_T} &\leq 2\left\|\sum_{k=1}^s \alpha_k(z)\phi_T(\theta^\star_k)\right\|^2_{L_T} + 2\left\|\sum_{k=1}^s \xi_k(z)\,\phi_T^{[1]}(\theta^\star_k)\right\|^2_{L_T} \\
&= 2\sum_{1\leq k,\ell\leq s} \mathcal{K}_T(\theta^\star_k, \theta^\star_\ell)\int \alpha_k(z)\alpha_\ell(z)\nu(\mathrm{d}z) + 2\sum_{1\leq k,\ell\leq s} \mathcal{K}_T^{[1,1]}(\theta^\star_k, \theta^\star_\ell)\int \xi_k(z)\xi_\ell(z)\nu(\mathrm{d}z) \\
&\leq 2\|\alpha\|_{*,q}\|\alpha\|_{*,p}\sum_{1\leq k,\ell\leq s} |\mathcal{K}_T(\theta^\star_k, \theta^\star_\ell)| + 2\|\xi\|_{*,q}\|\xi\|_{*,p}\sum_{1\leq k,\ell\leq s} |\mathcal{K}_T^{[1,1]}(\theta^\star_k, \theta^\star_\ell)| \\
&\leq 2s\,\|\alpha\|_{*,q}\|\alpha\|_{*,p}\left\|\Gamma^{[0,0]}\right\|_{\mathrm{op},*} + 2s\,\|\xi\|_{*,q}\|\xi\|_{*,p}\left\|\Gamma^{[1,1]}\right\|_{\mathrm{op},*}.
\end{aligned}$$

Using that $\|I\|_{\mathrm{op},*} = 1$ and (3.56), we get that:

$$\left\|\Gamma^{[0,0]}\right\|_{\mathrm{op},*} \leq 1 + u_T(s) \quad \text{and} \quad \left\|\Gamma^{[1,1]}\right\|_{\mathrm{op},*} \leq 1 + u_T(s).$$

By assumption we have $u_T(s) \leq H^{(2)}_\infty(r,\rho) \leq \frac{1}{6}$. Using (3.61), we deduce that:

$$\|P_{\alpha,\xi}\|^2_{L_T} \leq 2(1+u_T(s))\frac{(1-u_T(s))^2 + u_T(s)^2}{(1-2u_T(s))^2}\nu(\mathcal{Z})^{1/p-1/q}s \leq 4\,s\,\nu(\mathcal{Z})^{1/p-1/q}.$$

This gives:

$$\|P_{\alpha,\xi}\|_{L_T} \leq 2\sqrt{s}\,\nu(\mathcal{Z})^{1/2p-1/2q}. \tag{3.71}$$

We proved that (*i*)-(*iv*) from Assumption 3.4.1 stand. By assumption we also have that for all $\theta \neq \theta' \in \mathcal{Q}^\star : \mathfrak{d}_T(\theta,\theta') > 2\,r$, therefore Assumption 3.4.1 holds. This finishes the proof of Proposition 3.4.1.

3.8.2 Proof of Proposition 3.4.2 (Construction of an interpolating derivative certificate)

This section is devoted to the proof of Proposition 3.4.2. We shall closely follow the proof of [Butucea et al., 2022a, Proposition 7.5].

Let $T \in \mathbb{N}$ and $s \in \mathbb{N}^*$. Recall Assumptions 3.2.2 (and thus 3.2.1 on the regularity of φ_T) and 3.2.3 on the regularity of the asymptotic kernel $\mathcal{K}_\infty$ are in force. Let $r > 0$ and $u'_\infty \in (0, 1/6)$ such that (*iii*) and (*iv*) of Proposition 3.4.2 hold. Recall the definitions (3.23) and (3.24) of $\Theta^s_{T,\delta}$ and δ_∞. By assumption $\delta_\infty(u'_\infty, s)$ is finite. Let $\vartheta^\star = (\theta^\star_1, \dots, \theta^\star_s) \in \Theta^s_{T,(2\rho_T\,\delta_\infty(u'_\infty,s))\vee(2r)}$. We note $\mathcal{Q}^\star = \{\theta^\star_i,\, 1 \leq i \leq s\}$ the set of cardinal s.

Let $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ be such that $\|V(\theta^\star)\|_{L^q(\nu)} = 1$ for any $\theta^\star \in \mathcal{Q}^\star$. Recall the notation $\overline{V}$ defined in (3.57). Let $\alpha, \xi \in L^q(\nu, \mathbb{R}^s)$. We consider the real-valued function $\eta_{\alpha,\xi}$ defined on $\mathcal{Z} \times \Theta$ by (3.51).

Recall the definition of $\mathcal{V}_T$ from (3.14) and define $u'_T(s) = u'_\infty + (s-1)\mathcal{V}_T$. Thanks to (3.54) and (3.55), we get that (3.56) holds with $u_T(s)$ replaced by $u'_T(s)$.

Lemma 3.8.5. *Let be $1 \leq p \leq q \leq +\infty$ such that $1/p + 1/q = 1$. Let $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ be a measurable application such that for any $\theta^\star \in \mathcal{Q}^\star$, $\|V(\cdot, \theta^\star)\|_{L^q(\nu)} = 1$. Assume that we have (3.56) with $u_T(s)$ replaced by $u'_T(s) < 1/2$. Then, there exist $\alpha, \xi \in L^q(\nu, \mathbb{R}^s)$ such that:*

$$\eta_{\alpha,\xi}(z, \theta^\star_k) = 0 \quad \textit{for} \quad 1 \leq k \leq s, \ \textit{for } \nu - \textit{almost every } z, \tag{3.72}$$

$$\tilde{D}_{1,T}[\eta_{\alpha,\xi}(z)](\theta^\star_k) = V(z, \theta^\star_k) \quad \textit{for} \quad 1 \leq k \leq s, \ \textit{for } \nu - \textit{almost every } z, \tag{3.73}$$

and we also have:

$$\|\alpha\|_{*,q} \leq \frac{u_T'(s)}{1-2u_T'(s)}, \quad \|\xi\|_{*,q} \leq \frac{1-u_T'(s)}{1-2u_T'(s)}, \tag{3.74}$$

and

$$\|\alpha\|_{*,p} \leq \nu(\mathcal{Z})^{1/p-1/q}\frac{u_T'(s)}{1-2u_T'(s)}, \quad \|\xi\|_{*,p} \leq \nu(\mathcal{Z})^{1/p-1/q}\frac{1-u_T'(s)}{1-2u_T'(s)}. \tag{3.75}$$

Proof. Let $z \in \mathcal{Z}$ such that (3.72) and (3.73) are satisfied. Using the notations from Section 3.8.1, we obtain by [Butucea et al., 2022a, Lemma 10.2] that:

$$\alpha(z) = -\Gamma_{SC}^{-1}\Gamma^{[1,0]\top}[\Gamma^{[1,1]}]^{-1}\overline{V}(z) \quad \text{and} \quad \xi(z) = \left(I + [\Gamma^{[1,1]}]^{-1}\Gamma^{[1,0]}\Gamma_{SC}^{-1}\Gamma^{[1,0]\top}\right)[\Gamma^{[1,1]}]^{-1}\overline{V}(z).$$

Using (3.56), (3.60) and the fact that $\left\|\overline{V}\right\|_{*,q} = 1$, we readily obtain (3.74). We then obtain the controls (3.75) using (3.61).

□

We fix $V : \mathcal{Z} \times \mathcal{Q}^\star \to \mathbb{R}$ such that for any $\theta^\star \in \mathcal{Q}^\star$ we have $\|V(\theta^\star)\|_{L^q(\nu)} = 1$ and we consider $P_{\alpha,\xi}$ and $\eta_{\alpha,\xi}$ given by (3.50) and (3.51), with α and ξ given by Lemma 3.8.5.

Proof of *(i)* **from Assumption 3.4.2** with $c_N = (L_{0,2} + L_{2,1} + 7)/8$. We define the function $f : (z,\theta) \mapsto \eta_{\alpha,\xi}(z,\theta) - V(z,\theta_\ell^\star)\ \mathrm{sign}(\theta-\theta_\ell^\star)\mathfrak{d}_T(\theta,\theta_\ell^\star)$ on $\mathcal{Z}\times\Theta$. To prove the Property *(i)*, we will bound $\left\|\tilde{D}_{2;T}[f](\theta)\right\|_{L^q(\nu)}$ on Θ and apply Lemma 3.9.3-*(i)*. Recall $\mathfrak{d}_T(\theta,\theta_\ell^\star) = |G_T(\theta) - G_T(\theta_\ell^\star)|$ with G_T a primitive of $\sqrt{g_T}$, and thus $f(z,\theta) = \eta_{\alpha,\xi}(z,\theta) - V(z,\theta_\ell^\star)(G_T(\theta) - G_T(\theta_\ell^\star))$. We deduce that for ν-almost every $z \in \mathcal{Z}$ the function f is twice continuously differentiable with respect to its second variable on Θ; and elementary calculations give that $\tilde{D}_{2;T}[f(z)](\theta) = \tilde{D}_{2;T}[\eta_{\alpha,\xi}(z)](\theta)$ for any $\theta \in \Theta$ and for ν-almost every $z \in \mathcal{Z}$ as $\tilde{D}_{1;T}[G_T] = 1$ and $\tilde{D}_{2;T}[G_T] = 0$.

Let $\theta \in \Theta_T$ and let $\theta_\ell^\star$ be one of the elements of $\mathcal{Q}^\star$ closest to θ in terms of the metric $\mathfrak{d}_T$. Recall the notations $\Gamma_{\ell,\theta}$ (resp. $\Gamma_{\ell,\theta}^{[i,j]}$) and $\vartheta_{\ell,\theta}^\star$ defined after (3.62). Since for ν-almost every $z \in \mathcal{Z}$ we have $\tilde{D}_{2;T}[f(z)] = \tilde{D}_{2;T}[\eta_{\alpha,\xi}(z)]$, we deduce from (3.66) that:

$$\left\|\tilde{D}_{2;T}[f](\theta)\right\|_{L^q(\nu)} \leq \left\|I + \Gamma_{\ell,\theta}^{[2,0]}\right\|_{\mathrm{op},*}\|\alpha\|_{*,q} + \|\alpha\|_{*,q}|\mathcal{K}_T^{[2,0]}(\theta,\theta_\ell^\star)| + \|\xi\|_{*,q}\left\|\Gamma_{\ell,\theta}^{[2,1]}\right\|_{\mathrm{op},*} + \|\xi\|_{*,q}|\mathcal{K}_T^{[2,1]}(\theta,\theta_\ell^\star)|.$$

Notice that (3.62) holds with $u_T(s)$ replaced by $u_T'(s)$. Using (3.68) and (3.69) and the bounds (3.74) on α and ξ from Lemma 3.8.5, we get:

$$\left\|\tilde{D}_{2;T}[f](\theta)\right\|_{L^q(\nu)} \leq \frac{u_T'(s)}{1-2u_T'(s)}(L_{2,0} + \mathcal{V}_T + u_T'(s)) + \frac{1-u_T'(s)}{1-2u_T'(s)}(L_{2,1} + \mathcal{V}_T + u_T'(s)).$$

By assumption, we have $u_T'(s) \leq 1/6$ and $\mathcal{V}_T \leq 1$. Hence, we obtain:

$$\left\|\tilde{D}_{2;T}[f](\theta)\right\|_{L^q(\nu)} \leq \frac{1}{4}L_{2,0} + \frac{5}{4}L_{2,1} + \frac{7}{4}.$$

Since we have for almost every $z \in \mathcal{Z}$,

$$f(z,\theta_\ell^\star) = 0 \quad \text{and} \quad \tilde{D}_{1;T}[f(z)](\theta_\ell^\star) = \tilde{D}_{1;T}[\eta_{\alpha,\xi}(z)](\theta_\ell^\star) - V(z,\theta_\ell^\star) = 0,$$

using Lemma 3.9.3 *(i)*, we get, with $c_N = (L_{2,0} + 5L_{2,1} + 7)/8$:

$$\|\eta_{\alpha,\xi}(\theta) - V(\theta_\ell^\star)\ \mathrm{sign}(\theta-\theta_\ell^\star)\,\mathfrak{d}_T(\theta,\theta_\ell^\star)\|_{L^q(\nu)} = \|f(\theta)\|_{L^q(\nu)} \leq c_N\,\mathfrak{d}_T(\theta,\theta_\ell^\star)^2.$$

Proof of *(ii)* **from Assumption 3.4.2** with $c_F = (5L_{1,0} + 7)/4$. Let $\theta \in \Theta_T$, we shall prove that $\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq c_F$. Let $\theta_\ell^\star$ be one of the elements of $\mathcal{Q}^\star$ closest to θ in terms of the metric $\mathfrak{d}_T$. We deduce from (3.64) on the upper bound of $\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)}$, using (3.56), the inequality from (3.11), (3.65) and the bounds (3.74) on α and ξ from Lemma 3.8.5 that:

$$\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq \frac{u_T'(s)}{1-2u_T'(s)}\left(1+u_T'(s)\right) + \frac{1-u_T'(s)}{1-2u_T'(s)}\left(L_{1,0}+\mathcal{V}_T+u_T'(s)\right).$$

Since $u_T'(s) \leq 1/6$ and $\mathcal{V}_T \leq 1$, we obtain:

$$\|\eta_{\alpha,\xi}(\theta)\|_{L^q(\nu)} \leq \frac{5}{4}L_{1,0} + \frac{7}{4}\cdot$$

Proof of *(iii)* **from Assumption 3.4.2** with $c_B = 2$. Using very similar arguments as in the proof of (3.71) (taking care that the upper bound of the norms $\|\cdot\|_{*,q}$ and $\|\cdot\|_{*,p}$ of α and ξ are given by (3.74) and (3.75)) we also get $\|P_{\alpha,\xi}\|_{L_T} \leq 2\sqrt{s}\,\nu(\mathcal{Z})^{1/2p-1/2q}$.

We proved that *(i)*-*(ii)* from Assumption 3.4.2 stand for any $\theta \in \Theta_T$. Hence Assumption 3.4.2 holds for any positive r such that for all $\theta \neq \theta' \in \mathcal{Q}^\star : \mathfrak{d}_T(\theta,\theta') > 2\,r$. This finishes the proof of Proposition 3.4.2.

3.9 Auxiliary Lemmas

In this section, we provide the proofs of the intermediate results.

3.9.1 Proof of Proposition 3.1.3

We prove the optimization problem (3.3) is well posed. Denote the objective function of (3.3) by $F(B,\vartheta)$, that is the penalized risk. Then, we have:

$$\inf_{B\in L^2(\nu,\mathbb{R}^K),\vartheta\in\Theta_T^K} \quad F(B,\vartheta) \leq F(0,\vartheta^\star) = \frac{1}{2\nu(\mathcal{Z})}\|Y\|_{L_T}^2.$$

By Minkowski inequality, we have that $\|\cdot\|_{L^p(\nu,\mathbb{R}^K)} \leq \|\cdot\|_{\ell_1,L^p(\nu)}$. Indeed, we have for any $B \in L^2(\nu,\mathbb{R}^K)$:

$$\|B\|_{L^p(\nu,\mathbb{R}^K)} := \left\|\left(\sum_{k=1}^K B_k^2\right)^{\frac{1}{2}}\right\|_{L^p(\nu)} \leq \left\|\sum_{k=1}^K |B_k|\right\|_{L^p(\nu)} \leq \|B\|_{\ell_1,L^p(\nu)}.$$

Therefore, the minimization of F over B can be restricted to the centered closed ball $\mathcal{B}_0$ in $L^p(\nu,\mathbb{R}^K)$ of radius $\|Y\|_{L_T}^2/(\kappa 2\nu(\mathcal{Z}))$. We recall that the space $L^p(\nu,\mathbb{R}^K)$ is a reflexive Banach space whose dual is $L^q(\nu,\mathbb{R}^K)$ with $1/p+1/q=1$, see [Diestel and Uhl, 1977, Theorem 1 p.98]. By Kakutani Theorem, the closed balls of $L^p(\nu,\mathbb{R}^K)$ are therefore compact with respect to the weak topology, see [Brezis, 2011, Theorem 3.17]. In particular (3.3) amounts to minimizing F over the compact set $\mathcal{B}_0 \times \Theta_T^K$.

We show that the objective function is lower semi-continuous (lsc). Recall that a convex strongly continuous (that is, continuous with respect to the strong topology) real valued function defined on a Banach space is weakly lsc (that is, lsc with respect to the weak topology), see [Brezis, 2011, Corollary 3.9]. For any $B \in L^2(\nu,\mathbb{R}^K)$, we have:

$$\|B\|_{\ell_1,L^p(\nu)} \leq K^{\frac{1}{q}}\left(\sum_{k=1}^K \|B_k\|_{L^p(\nu)}^p\right)^{\frac{1}{p}} = K^{\frac{1}{q}}\left(\int \|B(z)\|_{\ell_p}^p \nu(\mathrm{d}z)\right)^{\frac{1}{p}}$$
$$\leq K^{\frac{1}{2}}\left(\int \|B(z)\|_{\ell_2}^p \nu(\mathrm{d}z)\right)^{\frac{1}{p}} = K^{\frac{1}{2}}\cdot\|B\|_{L^p(\nu,\mathbb{R}^K)},$$

where we used Hölder's inequality. We deduce that the function $B \mapsto \|B\|_{\ell_1, L^p(\nu)}$ is strongly continuous. Since it is also convex, we get it is weakly lsc.

Recall the space $(L_T, \|\cdot\|_{L_T})$ is a Hilbert space, see [Diestel and Uhl, 1977, Section IV]. The function $X \mapsto \|Y - X\|_{L_T}$ defined on L_T is weakly lsc as it is strongly continuous and convex. Then, since the function $\vartheta \mapsto \Phi(\vartheta)$ is continuous, we deduce that the function $(B, \vartheta) \mapsto B\Phi(\vartheta)$ is continuous from $L^p(\nu, \mathbb{R}^K) \times \mathbb{R}^K$ to L_T with respect to the product topology of the weak topology on $L^p(\nu, \mathbb{R}^K)$ and the usual topology on $\mathbb{R}^K$. Since the composition of a continuous function by a lsc function is a lsc function, we deduce that the function $(B, \vartheta) \mapsto \|Y - B\Phi(\vartheta)\|_{L_T}$ is lsc (with respect to product topology of the weak topology on $L^p(\nu, \mathbb{R}^K)$ and the usual topology on $\mathbb{R}^K$).

In conclusion, the objective function $(B, \vartheta) \mapsto F(B, \vartheta)$ is lsc (with respect to the product topology of the weak topology on $L^p(\nu, \mathbb{R}^K)$ and the usual topology on $\mathbb{R}^K$). Then, we conclude using that a lsc function on a compact set attains a minimum value, see [Aliprantis and Border, 2006, Theorem 2.43].

3.9.2 Tail bound for suprema of χ^2 processes

We give a tail bound for suprema of weighted χ^2 processes indexed on an interval $I \subset \mathbb{R}$.

Lemma 3.9.1. *Let $I \subset \mathbb{R}$ be a bounded interval. Assume that $X = (X(\theta), \theta \in I)$ is a real centered Gaussian process with Lipschitz sample paths. Consider the process $Y = \sum_{i=1}^{n} X_i^2$ where $(X_i,\ 1 \le i \le n)$ are independent copies of X. Then, for an arbitrary $\theta_0 \in I$ and for all $u > n \sup_{\theta \in I} \mathrm{Var}(X(\theta))$, we have:*

$$\mathbb{P}\left(\sup_I Y > u\right) \le \mathrm{e}^{-\frac{u}{\mathrm{Var}X(\theta_0)}\left(1-2\sqrt{\frac{n\mathrm{Var}X(\theta_0)}{u}}\right)} + 4\int_I \frac{\sqrt{\mathrm{Var}(X'(\theta))}}{2^{n/2}\Gamma(n/2)\sqrt{u}}\left(\frac{u}{\mathrm{Var}(X(\theta))}\right)^{(n+1)/2} \mathrm{e}^{-\frac{u}{2\mathrm{Var}(X(\theta))}}\,\mathrm{d}\theta. \quad (3.76)$$

Proof. Recall that I is a bounded interval. Hence, the process Y defined on I has Lipschitz sample paths. Then, applying Inequality (122) from [Butucea et al., 2022a] to the process Y and taking the expectation, we get, with $M = \sup_I Y$, $a = u > 0$, $b = u + \varepsilon$, $\varepsilon > 0$ and $x_0 = \theta_0$:

$$\int_u^{u+\varepsilon} \mathbb{P}(M \ge t)\,\mathrm{d}t \le \varepsilon \mathbb{P}(Y(\theta_0) \ge u) + \int_I \mathbb{E}\left[|Y'(\theta)|\mathbf{1}_{\{u<Y(\theta)<u+\varepsilon\}}\right]\mathrm{d}\theta. \quad (3.77)$$

The random variable $Y(\theta_0)$ is a standard χ^2 variable of degree n and therefore we have by [Obozinski et al., 2011, Lemma 11] for $u > n\mathrm{Var}(X(\theta_0))$:

$$\mathbb{P}\left(Y(\theta_0) \ge u\right) \le \mathrm{e}^{-\frac{u}{\mathrm{Var}(X(\theta_0))}\left(1-2\sqrt{\frac{n\mathrm{Var}(X(\theta_0))}{u}}\right)}. \quad (3.78)$$

Notice that (3.78) trivially holds if $\mathrm{Var}(X(\theta_0)) = 0$ as $u > 0$.

We now give a bound of the second term in the right-hand side of (3.77). Since (X_i', X_i) are independent Gaussian processes for $i = 1, \cdots, n$, we can write for a given $\theta \in I$:

$$X_i'(\theta) = \alpha_\theta X_i(\theta) + \beta_\theta G_i,$$

where $(G_i,\ 1 \le i \ \le n)$ are independent standard Gaussian random variables independent of the variables $(X_i(\theta),\ 1 \le i \le n)$ and:

$$\alpha_\theta = \frac{\mathbb{E}[X'(\theta)X(\theta)]}{\mathrm{Var}(X(\theta))} \quad \text{and} \quad \beta_\theta^2 = \mathrm{Var}(X'(\theta)) - \alpha_\theta^2 \mathrm{Var}(X(\theta)),$$

with the convention that $\alpha_\theta = 0$ if $\mathrm{Var}(X(\theta)) = 0$. Since $Y' = 2\sum_{i=1}^{n} X_i' X_i$ *a.e.*, we get that:

$$\mathbb{E}\left[|Y'(\theta)|\mathbf{1}_{\{u<Y(\theta)<u+\varepsilon\}}\right] \leq 2|\alpha_\theta|\mathbb{E}\left[Y(\theta)\mathbf{1}_{\{u<Y(\theta)<u+\varepsilon\}}\right] + 2|\beta_\theta|\mathbb{E}\left[\left|\sum_{i=1}^{n} X_i(\theta)G_i\right|\mathbf{1}_{\{u<Y(\theta)<u+\varepsilon\}}\right].$$

Since the variables $(G_i,\, 1 \leq i \leq n)$ and $(X_i(\theta),\, 1 \leq i \leq n)$ are independent, the variable $Z = \sum_{i=1}^{n} X_i(\theta)G_i$ contidionally to the variables $(X_i(\theta),\, 1 \leq i \leq n)$ is a standard Gaussian random variable of variance $Y(\theta)$. This implies that:

$$\mathbb{E}\left[\left|\sum_{i=1}^{n} X_i(\theta)G_i\right|\mathbf{1}_{\{u<Y(\theta)<u+\varepsilon\}}\right] = \sqrt{\frac{2}{\pi}}\,\mathbb{E}\left[\sqrt{Y(\theta)}\mathbf{1}_{\{u<Y(\theta)<u+\varepsilon\}}\right].$$

We deduce that:

$$\mathbb{E}\left[|Y'(\theta)|\mathbf{1}_{\{u<Y(\theta)<u+\varepsilon\}}\right] \leq 2\left(|\alpha_\theta|(u+\varepsilon) + \sqrt{\frac{2}{\pi}}\,|\beta_\theta|\sqrt{u+\varepsilon}\right)\mathbb{P}(u < Y(\theta) < u+\varepsilon),$$

The random variable $Y(\theta)$ is distributed as a χ^2 variable and has a density:

$$p_{Y(\theta)}(u) = \frac{u^{n/2-1}}{2^{n/2}\Gamma(n/2)}\left(\frac{1}{\mathrm{Var}(X(\theta))}\right)^{n/2} \mathrm{e}^{-\frac{u}{2\mathrm{Var}(X(\theta))}},$$

where by convention $p_{Y(\theta)}(u)$ is taken equal to 0 if $\mathrm{Var}(X(\theta)) = 0$ and where Γ denotes the gamma function.

Letting ε goes to 0 in (3.77), using (3.78), the right continuity of the cdf of M and the monotonicity of the density $p_{Y(\theta)}(u)$ of $Y(\theta)$ on $[n\,\mathrm{Var}X(\theta), +\infty[$, we deduce that for $u > n\,\sup_{\theta\in I}\mathrm{Var}(X(\theta))$:

$$\mathbb{P}(M \geq u) \leq \mathrm{e}^{-\frac{u}{\mathrm{Var}X(\theta_0)}\left(1-2\sqrt{\frac{n\mathrm{Var}X(\theta_0)}{u}}\right)} + 2\int_I \left(|\alpha_\theta|u + \sqrt{\frac{2}{\pi}}\,|\beta_\theta|\sqrt{u}\right) p_{Y(\theta)}(u)\,\mathrm{d}\theta. \tag{3.79}$$

We now bound the second term of the right-hand side of (3.79) in two steps. Using that $\beta_\theta^2 \leq \mathrm{Var}(X'(\theta))$, we get that:

$$\sqrt{\frac{2}{\pi}}\,|\beta_\theta|\,\sqrt{u}\,p_{Y(\theta)}(u) \leq \frac{1}{\sqrt{\pi}}\,\frac{\sqrt{\mathrm{Var}(X'(\theta))}}{2^{(n-1)/2}\Gamma(n/2)\sqrt{u}}\left(\frac{u}{\mathrm{Var}(X(\theta))}\right)^{n/2} \mathrm{e}^{-\frac{u}{2\mathrm{Var}(X(\theta))}}. \tag{3.80}$$

Thanks to the Cauchy-Schwarz inequality, we get $|\alpha_\theta| \leq \sqrt{\mathrm{Var}(X'(\theta))}/\sqrt{\mathrm{Var}(X(\theta))}$. We get that:

$$|\alpha_\theta|u\,p_{Y(\theta)}(u) \leq \frac{\sqrt{\mathrm{Var}(X'(\theta))}}{2^{n/2}\Gamma(n/2)\,\sqrt{u}}\left(\frac{u}{\mathrm{Var}(X(\theta))}\right)^{(n+1)/2} \mathrm{e}^{-\frac{u}{2\mathrm{Var}(X(\theta))}}. \tag{3.81}$$

Notice that (3.80) and (3.81) hold also if $\mathrm{Var}(X(\theta)) = 0$. Using that $\sqrt{\frac{2}{\pi}} + 1 \simeq 1.8 \leq 2$ and that $u \geq \sup_{\theta\in I}\mathrm{Var}(X(\theta))$, we deduce (3.76) from (3.79), (3.80) and (3.81). □

Recall the functions f_n and g_n defined by (3.43).

Lemma 3.9.2. *Let $T \in \mathbb{N}$ and $n \in \mathbb{N}^*$ be fixed. Let be $\mathcal{Z} = \{1, \cdots, n\}$. Suppose that Assumptions 3.2.1 and 3.2.2 hold. Let h be a function of class $\mathcal{C}^1$ from Θ_T to H_T, with Θ_T a sub-interval of Θ. Assume there exist finite constants C_1 and C_2 such that for all $\theta \in \Theta_T$:*

$$\|h(\theta)\|_T \leq C_1 \quad \textit{and} \quad \left\|\tilde{D}_{1;T}[h](\theta)\right\|_T \leq C_2.$$

Let $(W_T(z), z \in \mathcal{Z})$ be H_T-valued noise processes such that Assumption 3.3.1 holds. Let $a = (a_1, \cdots, a_n)$ be a sequence of nonnegative real numbers.
Set for any z in the set $\mathcal{Z}$ of cardinal n, $X(z) = (X(z,\theta) = \langle h(\theta), W_T(z)\rangle_T, \theta \in \Theta)$ and $Y = \sum_{z\in\mathcal{Z}} a_z X(z)^2$. Then, we have for $u \geq (n+1)\,\|a\|_{\ell_\infty}\sigma^2\Delta_T C_1^2$:

$$\mathbb{P}\left(\sup_{\theta\in\Theta_T} Y(\theta) > u\right) \leq f_n\left(\frac{u}{\sigma^2\|a\|_{\ell_\infty}\Delta_T C_1^2}\right) + \frac{4\,C_2\,|\Theta_T|_{\mathfrak{d}_T}}{C_1\,2^{n/2}} g_n\left(\frac{u}{\sigma^2\|a\|_{\ell_\infty}\Delta_T C_1^2}\right), \tag{3.82}$$

where $|\Theta_T|_{\mathfrak{d}_T}$ denotes the diameter of the interval Θ_T with respect to the metric $\mathfrak{d}_T$, $\|a\|_{\ell_\infty} = \max_{z\in\mathcal{Z}}|a_z|$ and Γ denotes the classical gamma function.

Proof. First we notice that:

$$\mathbb{P}\left(\sup_{\Theta_T} Y > u\right) \leq \mathbb{P}\left(\sup_{\Theta_T} Z > u/\|a\|_{\ell_\infty}\right), \tag{3.83}$$

where $Z = \sum_{z\in\mathcal{Z}} X(z)^2$. We shall apply Lemma 3.9.1 to the process Z.

Recall that the Gaussian processes $X(z)$ with $z \in \mathcal{Z}$ are independent with the same distribution as a process denoted $X = (X(\theta),\, \theta \in \Theta_T)$. The process X has Lipschitz sample paths on Θ_T and $X'(\theta) = \langle \partial_\theta h(\theta), w_T\rangle_T$ for *a.e.* $\theta \in \Theta_T$. By Assumption 3.3.1, we have for all $\theta \in \Theta_T$ and $z \in \mathcal{Z}$:

$$\mathrm{Var}(X(\theta)) \leq \sigma^2\Delta_T\|h(\theta)\|_T^2 \leq \sigma^2\Delta_T C_1^2. \tag{3.84}$$

We first consider the case where $\Theta_T = [\theta_{\min}, \theta_{\max}]$ is a compact interval with $\theta_{\min} < \theta_{\max}$. Then, according to Lemma 3.9.1, Inequality (3.76) holds with Y replaced by Z for $u > n\,\sigma^2\Delta_T C_1^2$.

Notice that the function $x \mapsto x^{\frac{n+1}{2}}\,\mathrm{e}^{-x/2}$ is decreasing on $[n+1, +\infty)$ and that the function $x \mapsto \mathrm{e}^{-x\left(1-2\sqrt{\frac{n}{x}}\right)}$ is decreasing on $[n, +\infty)$. Then, plugging (3.84) in Inequality (3.76), we obtain for $u > (n+1)\,\sigma^2\Delta_T C_1^2$:

$$\begin{aligned}\mathbb{P}\left(\sup_{\Theta_T} Z > u\right) \leq{}& \mathrm{e}^{-\frac{u}{\sigma^2\Delta_T C_1^2}\left(1-2\sqrt{\frac{n\sigma^2\Delta_T C_1^2}{u}}\right)} \\ &+ \frac{4}{2^{n/2}\Gamma(n/2)\,\sqrt{u}}\left(\frac{u}{\sigma^2\Delta_T C_1^2}\right)^{(n+1)/2}\mathrm{e}^{-\frac{u}{2\sigma^2\Delta_T C_1^2}}\int_{\Theta_T}\sqrt{\mathrm{Var}(X'(\theta))}\,\mathrm{d}\theta.\end{aligned} \tag{3.85}$$

There exists a geodesic $\gamma : [0,1] \mapsto \Theta_T$ such that $\gamma_0 = \theta_{\min}$, $\gamma_1 = \theta_{\max}$ and $\mathfrak{d}_T(\theta_{\min}, \theta_{\max}) = \int_0^1 |\dot\gamma_t|\,\sqrt{g_T(\gamma_t)}\,\mathrm{d}t$. Hence, a change of variable gives:

$$\int_{\Theta_T}\sqrt{\mathrm{Var}(X'(\theta))}\,\mathrm{d}\theta = \int_0^1 |\dot\gamma_t|\sqrt{g_T(\gamma_t)\cdot\frac{\mathrm{Var}(X'(\gamma_t))}{g_T(\gamma_t)}}\,\mathrm{d}t. \tag{3.86}$$

By Assumption 3.3.1, we have for all $\theta \in \Theta_T$:

$$\frac{\mathrm{Var}(X'(\theta)))}{g_T(\theta)} \leq \sigma^2\Delta_T\left\|\tilde D_{1;T}[h](\theta)\right\|_T^2 \leq \sigma^2\Delta_T C_2^2.$$

Using this bound in (3.86), we get:

$$\int_{\Theta_T}\sqrt{\mathrm{Var}(X'(\theta))}\,\mathrm{d}\theta \leq C_2\,\sigma\,\sqrt{\Delta_T}\,|\Theta_T|_{\mathfrak{d}_T}, \tag{3.87}$$

where $|\Theta_T|_{\mathfrak{d}_T}$ is the diameter of the interval Θ_T with respect to the metric $\mathfrak{d}_T$.

Combining (3.85), (3.87) and (3.83), we finally obtain (3.82) for Θ_T a bounded closed interval. Then, use monotone convergence and the continuity of Z to get (3.82) for any interval Θ_T. □

3.9.3 Technical lemma

We consider functions $\eta : \mathcal{Z} \times \Theta \mapsto \mathbb{R}$ and bound the quantities $\|\eta(\theta)\|_{L^q(\nu)}$ on some regions of Θ under some assumptions on the second covariant derivative of η with respect to θ. The following Lemma extends [Poon et al., 2021, Lemma 2]. The proof is similar, as the latter covers the case where ν is a Dirac measure and $\|\cdot\|_{L^q(\nu)}$ reduces to $|\cdot|$.

Lemma 3.9.3. *Let $q \in [1, +\infty]$. Suppose Assumption 3.2.2 holds. Consider a function $\eta : \mathcal{Z} \times \Theta$ twice continuously differentiable with respect to its second variable and $\theta_0 \in \Theta_T$.*

(i) Assume that for ν-almost every $z \in \mathcal{Z}$ we have $\eta(z, \theta_0) = 0$ and $\tilde{D}_{1;T}[\eta(z)](\theta_0) = 0$, and that there exist $\delta > 0$ and $r > 0$ such that for any $\theta \in \mathcal{B}_T(\theta_0, r)$ we have:

$$\left\|\tilde{D}_{2;T}[\eta](\theta)\right\|_{L^q(\nu)} \leq \delta.$$

Then, we have $\|\eta(\theta)\|_{L^q(\nu)} < (\delta/2)\, \mathfrak{d}_T(\theta, \theta_0)^2$, for any $\theta \in \mathcal{B}_T(\theta_0, r)$.

(ii) Assume now that for ν-almost every $z \in \mathcal{Z}$, $\eta(z, \theta_0) = V(z)$ and $\tilde{D}_{1;T}[\eta(z)](\theta_0) = 0$ where $V \in L^q(\nu)$ with $\|V\|_{L^q(\nu)} = 1$. Assume there exists a finite positive constant L such that $\sup_{\theta_0, \theta \in \Theta_T} |\mathcal{K}_T^{[0,2]}(\theta_0, \theta)| \leq L$ and there exist $\varepsilon > 0$ and $r \in (0, L^{-\frac{1}{2}})$ such that for any $\theta \in \mathcal{B}_T(\theta_0, r)$, $-\mathcal{K}_T^{[0,2]}(\theta_0, \theta) \geq \varepsilon$. Suppose that for any $\theta \in \mathcal{B}_T(\theta_0, r)$ and $\delta < \varepsilon$:

$$\left\|\tilde{D}_{2;T}[\eta](\theta) - V\mathcal{K}_T^{[0,2]}(\theta_0, \theta)\right\|_{L^q(\nu)} \leq \delta.$$

Then, we have $\|\eta(\theta)\|_{L^q(\nu)} \leq 1 - \frac{(\varepsilon - \delta)}{2} \mathfrak{d}_T(\theta, \theta_0)^2$, for any $\theta \in \mathcal{B}_T(\theta_0, r)$.

4 Off-the-grid prediction and testing for mixtures of translated features

Contents

Preamble

We consider a model where a signal (discrete or continuous) is observed with an additive Gaussian noise process. The signal is issued from a linear combination of a finite but increasing number of translated features. The features are continuously parameterized by their location and depend on some scale parameter. First, we extend previous prediction results for off-the-grid estimators by taking into account here that the scale parameter may vary. The prediction bounds are analogous, but we improve the minimal distance between two consecutive features locations in order to achieve these bounds.

Next, we propose a goodness-of-fit test for the model and give non-asymptotic upper bounds of the testing risk and of the minimax separation rate between two distinguishable signals. In particular, our test encompasses the signal detection framework. We deduce upper bounds on the minimal energy, expressed as the ℓ_2-norm of the linear coefficients, to successfully detect a signal in presence of noise. The general model considered in this chapter is a non-linear extension of the classical high-dimensional regression model. It turns out that, in this framework, our upper bound on the minimax separation rate matches (up to a logarithmic factor) the lower bound on the minimax separation rate for signal detection in the high dimensional linear model associated to a fixed dictionary of features. We also propose a procedure to test whether the features of the observed signal belong to a given finite collection under the assumption that the linear coefficients may vary, but do not change to opposite signs under the null hypothesis. A non-asymptotic bound on the testing risk is given.

We illustrate our results on the spikes deconvolution model with Gaussian features on the real line and with the Dirichlet kernel, frequently used in the compressed sensing literature, on the torus.

The material of this chapter has been released in [Butucea et al., 2022b].

4.1 Introduction

4.1.1 Model

This chapter is motivated by the study of the spikes deconvolution model [Duval and Peyré, 2015] with applications in spectroscopy ([Butucea et al., 2021]). In this model, a linear combination (or mixture) of spikes continuously parameterized is observed with an additive Gaussian noise process. We assume that the spikes are parameterized by a location parameter, that the noise and the observation space can vary with some parameter T increasing with the quality of the observations. More general non-linear models for the spikes have been discussed in [Butucea et al., 2022a], and the particular case of location families has been discussed in Section 8 of that paper. However, we allow here the scale of the spikes to vary with T, which makes the approach very different from the previous one.

We are also interested in goodness-of-fit testing, that is we want to test whether the observations are issued from a given linear combination of spikes. We remark that it includes the case of signal detection. This test problem finds an application in spectroscopy to detect the presence of a chemical compound in a material. In addition, we are interested in testing whether the observed signal is a linear combination of spikes located at a prescribed list of locations with linear coefficients having prescribed signs under the null hypothesis.

Let $T \in \mathbb{N}$. We observe a random element y in the Hilbert space $L^2(\lambda_T)$ of square integrable functions with respect to the measure λ_T on the Borel σ-field of some metric space Θ. The observation is the sum of a deterministic signal and a Gaussian random process w_T in $L^2(\lambda_T)$. We shall assume that the signal is an unknown finite mixture of s features belonging to a continuously parametrized subfamily $(\varphi_T(\theta),\, \theta \in \Theta)$ of $L^2(\lambda_T)$. We call this family a continuous dictionary, the weights of the mixture - the linear coefficients, and the parameters of the features - the non-linear parameters.

The quality of the information provided by the observations depends on the support of the measure λ_T and on the noise w_T. It increases with the parameter T. For example, we will consider the case where the sequence of measures $(\lambda_T, T \in \mathbb{N})$ converges towards the Lebesgue measure, noted Leb, on Θ, so that in the limit model the observation corresponds to a square integrable random process indexed on Θ. We consider the cases where $\Theta = \mathbb{R}$ and the limit measure is the Lebesgue measure on $\mathbb{R}$ as well as the case where Θ is the torus $\mathbb{R}/\mathbb{Z}$ and the limit measure is the Lebesgue measure on this manifold.

We consider in this chapter a dictionary given by the location model:

$$\Big(\varphi_T(\theta) = h(\theta - \cdot, \sigma_T),\, \theta \in \Theta\Big) \tag{4.1}$$

where h is a real-valued function defined on $\Theta \times \mathfrak{S}$, smooth with respect to its first variable and normalized so that $\|h(\cdot, \sigma_T)\|_{L^2(\mathrm{Leb})} = 1$, and where σ_T is an element of the set $\mathfrak{S}$ of admissible positive scale parameters. See Section 4.2.2 for examples of functions h including the Gaussian spike and the low-pass filter. Even though the location model considered here is a restriction when compared to general non-linear dictionnaries of features considered by e.g. [Butucea et al., 2022a], the scaling σ_T introduced here makes this dictionnary different. Indeed, this scaling is allowed to depend on T and may improve previous results in the sense that the sufficient conditions on the non-linear parameters in the mixture in order to obtain the prediction and estimation bounds are milder. The least separation distance between the location parameters in this model is allowed to be smaller when compared to unscaled dictionnaries, see Remark 4.2.2.

The Hilbert space $L^2(\lambda_T)$ is endowed with the natural scalar product noted $\langle \cdot, \cdot \rangle_{L^2(\lambda_T)}$ and norm $\|\cdot\|_{L^2(\lambda_T)}$. Let us define the normalized function ϕ_T defined on Θ by:

$$\phi_T(\theta) = \varphi_T(\theta) / \|\varphi_T(\theta)\|_{L^2(\lambda_T)},$$

and the multivariate function Φ_T on Θ^s by:

$$\Phi_T(\vartheta) = (\phi_T(\theta_1), \ldots, \phi_T(\theta_s))^\top \quad \text{for} \quad \vartheta = (\theta_1, \cdots, \theta_s) \in \Theta^s.$$

We assume that the signal contains an unknown number $s \in \mathbb{N}$ of active features. We consider the model with unknown non-zero linear coefficients $\beta^\star$ in $(\mathbb{R}^*)^s$ and unknown distinct parameters $\vartheta^\star = (\theta_1^\star, \cdots, \theta_s^\star) \in \Theta^s$:

$$y = \beta^\star \Phi_T(\vartheta^\star) + w_T \quad \text{in } L^2(\lambda_T), \tag{4.2}$$

where when $s = 0$, we set by convention that $\beta^\star \Phi_T(\vartheta^\star) = 0$ as well as $A^s = \{0\}$ for any set A. We denote by $\mathcal{Q}^\star = \{\theta_\ell^\star, 1 \leq \ell \leq s\}$ the set of the non-linear parameters associated to an active feature. The process y is observed over the support of the measure λ_T. Therefore it is legimate to consider models whose location parameters belong to the smallest interval covering the support of the measure λ_T. Hence, we introduce the set Θ_T, a compact interval of Θ, and we shall assume that $\mathcal{Q}^\star$ is a subset of Θ_T. We denote by $|\Theta_T|$ the Euclidean diameter of the set Θ_T.

We consider a large variety of Gaussian noise processes. Indeed, we only assume the following mild assumption on w_T, where the decay rate $\Delta_T > 0$ controls the noise variance decay as the parameter T grows and $\overline{\sigma} > 0$ is the intrinsic noise level. A wide range of noise processes satisfy our assumptions, see [Butucea et al., 2022a]; they can be discrete or continuous, white or coloured under these constraints.

Assumption 4.1.1 (Admissible noise)**.** *Let $T \in \mathbb{N}$. The Gaussian noise process w_T satisifies $\mathbb{E}\left[\|w_T\|^4_{L^2(\lambda_T)}\right] < +\infty$, and there exist a noise level $\overline{\sigma} > 0$ and a decay rate $\Delta_T > 0$ such that for all $f \in L^2(\lambda_T)$, the random variable $\langle f, w_T\rangle_{L^2(\lambda_T)}$ is a centered Gaussian random variable satisfying:*

$$\mathrm{Var}\left(\langle f, w_T\rangle_{L^2(\lambda_T)}\right) \leq \overline{\sigma}^2 \, \Delta_T \, \|f\|^2_{L^2(\lambda_T)}.$$

We assume that the quantity $\mathbb{E}\left[\|w_T\|^2_{L^2(\lambda_T)}\right]$ is known for the considered models. We consider the variance of the squared norm of the noise:

$$\Xi_T = \mathrm{Var}\left(\|w_T\|^2_{L^2(\lambda_T)}\right). \tag{4.3}$$

4.1.2 Examples of Gaussian noise processes

We consider a large variety of models: discrete models where the process y is observed on a grid or continuous models where the process is observed on an interval.

4.1.2.1 Discrete-time process observed on a regular grid

Consider a real-valued process y observed over a regular grid $t_1 < \cdots < t_T$ of a symmetric interval $[a_T, b_T] \subset \mathbb{R}$, with $T \geq 1$, $a_T = -b_T < 0$, $t_j = a_T + j\Delta_T$ for $j = 1, \ldots, T$ and grid step: $\Delta_T = (b_T - a_T)/T$. We set $\lambda_T = \Delta_T \sum_{j=1}^T \delta_{t_j}$ and see y as an element of $L^2(\lambda_T)$. We assume that $(b_T, T \geq 2)$ is a sequence of positive numbers, such that: $\lim_{T\to\infty} b_T = +\infty$ and $\lim_{T\to\infty} \Delta_T = 0$ so that the sequence of measures $(\lambda_T, T \geq 1)$ converges with respect to the vague topology towards the Lebesgue measure. In this formalism, the noise $w_T \in L^2(\lambda_T)$ is given by:

$$w_T(t) = \sum_{j=1}^T G_j \mathbf{1}_{\{t_j\}}(t), \tag{4.4}$$

where $\mathbf{1}_A$ denotes the indicator function of an arbitrary set A and $(G_1, \cdots, G_T)$ is a centered Gaussian random vector with independent entries of variance $\overline{\sigma}^2$. In this case, we have $\mathbb{E}[\|w_T\|^4_{L^2(\lambda_T)}] = \overline{\sigma}^4 \Delta_T^2 T(T+2)$ and Assumption 4.1.1 holds with an equality. We readily obtain that $\Xi_T = 2\overline{\sigma}^4 \Delta_T^2 T$.

In this particular example we have for any function $f \in L^2(\lambda_T)$ that $\|f\|_{L^2(\lambda_T)} = \sqrt{\Delta_T}\|f\|_{\ell_2}$, where the right-hand side is understood as the ℓ_2-norm (Euclidean norm) of the vector $(f(t_1), \ldots, f(t_T))$.

4.1.2.2 Continuous-time processes

Assume we observe a process y on an interval. We note λ_T for a σ-finite measure on $\mathbb{R}$ or on $\mathbb{R}/\mathbb{Z}$. In this framework, y is an element of $L^2(\lambda_T)$. Let us assume that the noise is $w_T = \sum_{k\in\mathbb{N}} \sqrt{\xi_k}\, G_k\, \psi_k$, where $(G_k, k \in \mathbb{N})$ are independent centered Gaussian random variables with variance $\overline{\sigma}^2$, $(\psi_k, k \in \mathbb{N})$ is an o.n.b. of $L^2(\lambda_T)$ on $\mathbb{R}$ or on $\mathbb{R}/\mathbb{Z}$, and that $\xi = (\xi_k, k \in \mathbb{N})$ is a square summable sequence of non-negative real numbers. We remark that Assumption 4.1.1 holds as $\mathbb{E}[\|w_T\|_T^4] = 3\overline{\sigma}^4 \sum_{k\in\mathbb{N}} \xi_k^2 + \overline{\sigma}^4 \sum_{k,\ell\in\mathbb{N},k\neq\ell} \xi_k\xi_\ell$ is finite and $\mathrm{Var}(\|w_T\|^2_{L^2(\lambda_T)}) = 2\overline{\sigma}^4 \sum_{k\in\mathbb{N}} \xi_k^2$. Moreover, we have:

$$\mathrm{Var}(\langle f, w_T\rangle_{L^2(\lambda_T)}) = \overline{\sigma}^2 \sum_{k\in\mathbb{N}} \xi_k \langle f, \psi_k\rangle^2_{L^2(\lambda_T)} \leq \overline{\sigma}^2\, \Delta_T\, \|f\|^2_{L^2(\lambda_T)} \quad \text{with} \quad \Delta_T = \sup_{k\in\mathbb{N}} \xi_k.$$

In this example the noise w_T depends on the parameter T only if ξ, and thus Δ_T, depend on T. We may consider different choices for ξ that lead to different values for Ξ_T, the variance of the squared norm of the noise. For instance, our framework encompasses the truncated white noise by taking for all $k \in \mathbb{N}$, $\xi_k = T^{-1}\mathbf{1}_{\{1\leq k\leq T\}}$. In this case, elementary calculations give $\Delta_T = 1/T$ and $\Xi_T = 2\overline{\sigma}^4/T$.

4.1.3 Description of the results

The aim of this chapter is twofold. First, we improve on [Butucea et al., 2022a] in the case of linear combination of translated spikes by giving bounds on the prediction error under milder separation constraints between the non-linear parameters in $\mathcal{Q}^\star$. This is achieved by taking the scale parameter of the features σ_T into account. In particular, in the case of Gaussian spikes deconvolution, the separation is of order σ_T.

Then, test problems are studied. We give procedures for the goodness-of-fit of the mixture model in order to determine whether the unknown signal $\beta^\star\Phi_T(\vartheta^\star)$ is equal to a reference signal $\beta^0\Phi_T(\vartheta^0)$ for some known vectors $\beta^0 \in (\mathbb{R}^*)^{s^0}$ and $\vartheta^0 \in \Theta_T^{s^0}$. This setup includes the case of signal detection where the null hypothesis is $\beta^\star \equiv 0$, that is $s = 0$. We propose a combined procedure based on differences between the reference signal $\beta^0\Phi_T(\vartheta^0)$ and either the observation y or a reconstructed signal obtained from estimators of the model parameters. In order to successfully perform the test, we remove from the alternative hypothesis the signals whose proximity with the reference signal $\beta^0\Phi_T(\vartheta^0)$ with respect to the norm $\|\cdot\|_{L^2(\lambda_T)}$ is below some separation parameter. We give a non-asymptotic upper bound of the testing risk and deduce an upper bound on the minimal separation needed to distinguish two different signals. This upper bound yields two regimes depending on whether the observed signal and the reference signal are sparse or not. In the case of signal detection, the separation can be expressed as the ℓ_2-norm of the linear coefficients of the observed mixture. In particular, when the observation y is issued from a non-linear extension of the classical high-dimensional regression model, our upper bound matches (up to logarithmic factors) the asymptotic lower bound of the minimal separation needed to distinguish two signals that are mixture of features from a finite high-dimensional dictionary.

We also test the presence of at most s_0 prescribed features in the mixture with arbitrary linear coefficients of given sign. That is, we test whether for each $\epsilon = \pm 1$ the unknown set $\mathcal{Q}^{\star,\epsilon} = \{\theta_k^\star \in \mathcal{Q}^\star : \epsilon\beta_k^\star > 0\}$ is a subset of $\mathcal{Q}^{0,\epsilon}$, with $\mathcal{Q}^{0,+}$ and $\mathcal{Q}^{0,-}$ being disjoint finite subsets of the set Θ_T. This setup is issued from an application to spectroscopy (see [Butucea et al., 2021]), where the presence of other chemical components than the prescribed ones are indicating aging or substantial modifications of the analyzed material. To separate the null hypothesis from the alternative hypothesis, we introduce a discrepancy that is 0 under all parameters $(\beta^\star, \vartheta^\star)$ belonging to the null hypothesis. We give an upper bound on the minimal separation to successfully perform our test. The test statistic introduced and studied in this context makes explicit use of the certificates used in [Butucea et al., 2022a] for establishing the prediction rates of the estimators of $(\beta^\star, \vartheta^\star)$. We stress the fact that

the test statistic is not an estimator of the discrepancy measure separating the null and the alternative hypotheses, as is usually the case in non-parametric tests.

4.1.4 Previous work

Estimating the linear coefficients and the parameters of model (4.2) from an observation y has attracted a lot of attention over the past decade. A major contribution in this field comes from the formulation of the BLasso problem in [de Castro and Gamboa, 2012]. This optimization problem on a space of measures allows to estimate both linear coefficients and non-linear parameters without using a grid on the parameter space. This off-the-grid method has successfully been used in [Candès and Fernandez-Granda, 2014] and [Candès and Fernandez-Granda, 2013] in the context of super-resolution as well as in [Duval and Peyré, 2015] for spikes deconvolution. High probability bounds for the prediction error have been given in [Tang et al., 2015] and [Boyer et al., 2017] for the specific dictionary of complex exponentials continuously paremetrized by their frequencies and more recently in [Butucea et al., 2022a] for a wide range of dictionnaries parametrized over a one-dimensional space. These results are based on certificate functions whose existence have been proven in a very general framework in [Poon et al., 2021] provided that the non-linear parameters of the mixture are well-separated with respect to a Riemannian metric.

Goodness-of-fit tests are used to check whether observations are indeed derived from a given statistical model. We refer to the monograph [Ingster and Suslina, 2003] for a comprehensive presentation of goodness-of-fit testing. When we consider a finite dictionary of features $(\varphi_T(\theta), \theta \in \mathcal{Q})$ with $\mathcal{Q}$ a known finite subset of Θ, the model (4.2) can be rewritten as a linear regression model, possibly of high dimension depending on the size of the finite dictionary $p := \mathrm{Card}(\mathcal{Q})$. In this case, testing the goodness-of-fit of the model amounts to testing whether the linear coefficients in the mixture are equal to some given linear coefficients. When the dictionary is known, the testing problem is homogeneous in the linear coefficients β and is therefore equivalent to testing $\beta \equiv 0$, which is a signal detection problem.

Signal detection has raised a lot of interest over the past decades. It is well known that the alternative hypothesis H_1 (presence of signal) must be well separated from the null hypothsesis H_0 (only noise) in order to have tests with small risks. The separation can be seen as a minimal signal intensity allowing the detection. Then, it is a matter of interest to evaluate the minimax separation rate, i.e., the smallest separation that allows to distinguish the tested hypotheses. In [Ermakov, 1990], asymptotic rates for the minimax separation in the framework of signal detection are derived for the non-parametric Gaussian white noise model. Non-asymptotic rates were then derived in [Baraud, 2002] and later in [Laurent et al., 2012] to tackle the case of heterogeneous variances. We refer to the monograph [Giné and Nickl, 2016] for an overview of non-parametric hypotheses testing. Regarding the high dimensional regression model where the observation is of dimension T and the dictionary is fixed, known and of size p, the work of [Ingster et al., 2010] established the following asymptotic minimax separation rates under coherence assumptions on the dictionary:

$$\frac{1}{T^{\frac{1}{4}}} \wedge \sqrt{\frac{s}{T}\log(p)} \wedge \frac{p^{\frac{1}{4}}}{\sqrt{T}}.$$

The signal intensity is expressed by the ℓ_2-norm of the linear coefficients. Their lower bounds on the asymptotic minimax separation stand for both fixed and random designs whereas their upper bounds stand for random designs. The work of [Arias-Castro et al., 2011] does not tackle the high dimension but provides tests achieving the minimax separation for fixed designs under coherence assumptions on the dictionary.

In this chapter we shall consider that our features come from a continuous dictionary and have unknown location parameters. Hence, the existing results do not apply. Furthermore, for the considered non-linear extension of linear regression models, goodness-of-fit testing

does not reduce to signal detection. Therefore, we introduce new testing procedures. We stress that one of the test statistics is not derived from estimators of the linear coefficients. In fact, depending on the sparsity of the signal, the dimension of the observation and the size of the dictionary, plug-in methods using sparse estimators might not be the best way to proceed. They do not always lead to the minimal separation. In this sense, testing is a very different statistical problem from estimation.

4.1.5 Roadmap of the chapter

In Section 4.2, we start by presenting the assumptions needed to perform a successful estimation of the linear coefficients and location parameters of our model. After giving a prediction bound in Theorem 4.2.3, we show in Lemma 4.2.4 that the required assumptions are sufficient conditions for the identifiability of the model. In Section 4.3, we test whether the observation derives from a given mixture or from some other mixture sufficiently separated from the latter. We give in Theorems 4.3.1 and 4.3.3 bounds of the testing risks associated to two different test procedures. We show in Corollaries 4.3.2 and 4.3.5 that these two tests give two regimes for our upper bound on the minimal separation to distinguish two different signals from an observation contaminated by noise. We also provide a discussion on the comparison of our upper bounds with some existing lower bounds. In Section 4.4, we propose a procedure to test whether the active features in the observed signal belong to a given finite collection with linear coefficients of prescribed signs. Both hypotheses of this test problem are composite and a new measure of the separation between these hypotheses has been introduced. The proposed test makes use of the certificates used in the proof of the prediction bounds in an original way. A bound of the testing risk is given in Theorem 4.4.3 and in Corollary 4.4.4, we give an upper bound on the minimax separation rate. The examples of Gaussian scaled spikes deconvolution on $\mathbb{R}$ and low-pass filter on $\mathbb{R}/\mathbb{Z}$ are adressed in Sections 4.5 and 4.6.

4.2 Assumptions and prediction bounds

We recall in this section assumptions and definitions from Sections 3-5 of [Butucea et al., 2022a] in a simpler way adapted to our framework. In [Butucea et al., 2022a], the authors established high probability bounds for prediction and estimation errors associated to some estimators of $\beta^\star$ and $\vartheta^\star$ tackling a wider range of dictionaries.

4.2.1 Regularity of the features

We gather in this section the hypotheses that will be required on the features defined by (4.1).

Recall that the parameter space Θ is either $\mathbb{R}$ or the torus $\mathbb{R}/\mathbb{Z}$ endowed with the Lebesgue measure Leb. For convenience, we write $|x-y|$ for the Euclidean distance between x and y either on $\mathbb{R}$ or on the torus. Recall also that $L^2(\lambda_T)$ and $L^2(\mathrm{Leb})$ are the sets of square integrable functions on Θ with respect to the measures λ_T and Leb respectively. We denote $\mathfrak{S}$ the set of scale parameters.

Assumption 4.2.1 (Smoothness of the features)**.** *Let h be a function defined on $\Theta \times \mathfrak{S}$. Let $T \in \mathbb{N}$ and $\sigma_T \in \mathfrak{S}$. We assume that the function $\theta \mapsto h(\theta, \sigma_T)$ is of class $\mathcal{C}^3$ on Θ. We assume furthermore that $\|h(\cdot, \sigma_T)\|_{L^2(\mathrm{Leb})} = 1$, and that for all $\theta \in \Theta$ $\|h(\theta - \cdot, \sigma_T)\|_{L^2(\lambda_T)} > 0$ and all $i \in \{0, \cdots, 3\}$ $\|\partial_\theta^i h(\cdot, \sigma_T)\|_{L^2(\mathrm{Leb})} < +\infty$ and $\|\partial_\theta^i h(\theta - \cdot, \sigma_T)\|_{L^2(\lambda_T)} < +\infty$.*

Recall the function φ_T defined by (4.1) and notice that Assumption 4.2.1 implies that $\|\varphi_T(\theta)\|_{L^2(\lambda_T)} > 0$ on Θ. We define the function:

$$g_T(\theta) = \|\partial_\theta \phi_T(\theta)\|^2_{L^2(\lambda_T)}, \quad \text{where } \phi_T(\theta) = \varphi_T(\theta)/\|\varphi_T(\theta)\|_{L^2(\lambda_T)}. \tag{4.5}$$

Assumption 4.2.2 (Positivity of g_T)**.** *Assumption 4.2.1 holds and we have $g_T > 0$ on Θ.*

Let us mention that if for all $\theta \in \Theta$, $\varphi_T(\theta)$ and $\partial_\theta \varphi_T(\theta)$ are linearly independent functions of $L^2(\lambda_T)$ and $\|\partial_\theta \varphi_T(\theta)\|_{L^2(\lambda_T)} > 0$, then $g_T(\theta) > 0$ for all $\theta \in \Theta$ (see [Butucea et al., 2022a, Lemma 3.1]).

4.2.2 Examples of feature functions

We provide some examples from the literature.

(i) *Spike deconvolution.* The noisy mixture of translated and scaled Gaussian features corresponds to:

$$h(t, \sigma) \mapsto \frac{\exp(-t^2/2\sigma^2)}{\pi^{1/4}\sigma^{1/2}} \quad \text{on } \Theta \times \mathfrak{S} = \mathbb{R} \times \mathbb{R}_+^*. \tag{4.6}$$

The example of Gaussian spikes deconvolution is analyzed in full details in [Butucea et al., 2022a, Section 8] when σ_T does not depend on T. We shall consider here that the scale parameter σ_T may vary with T.

(ii) *Multi-resolution approximation.* We consider the normalized Shannon scaling function:

$$h(t, \sigma) \mapsto \sqrt{\sigma}\, \frac{\sin(\pi t/\sigma)}{\pi t} \quad \text{on } \Theta \times \mathfrak{S} = \mathbb{R} \times \mathbb{R}_+^*.$$

The associated dictionary allows to recover functions whose Fourier transform have their support in $[-\pi/\sigma, \pi/\sigma]$ (see [Mallat, 2009, Theorem 3.5]).

(iii) *Low-pass filter.* We consider the normalized Dirichlet kernel on the torus for some cut-off frequency $f_c \in \mathbb{N}^*$ and $T = 2f_c + 1$:

$$h(t, \sigma) = \frac{1}{\sqrt{T}} \sum_{k=-f_c}^{f_c} \mathrm{e}^{2i\pi kt} = \frac{\sin(T\pi t)}{\sqrt{T}\,\sin(\pi t)}, \quad \text{with } \sigma = \frac{1}{T},\ T \in 2\mathbb{N}^* + 1 \text{ and } t \in \Theta = \mathbb{R}/\mathbb{Z}. \tag{4.7}$$

The example of the low-pass filter is adressed in [Duval and Peyré, 2015], where exact support recovery results are obtained for the BLasso estimators. This dictionary is also used in [Candès and Fernandez-Granda, 2013] in the context of super-resolution. Bounds on some prediction risks (different from those considered in this chapter) are established therein for estimators obtained by solving the constrained formulation of the BLasso.

4.2.3 Definition of the kernel and its approximation

4.2.3.1 Measuring the colinearity of the features

We define the symmetric kernel $\mathcal{K}_T$ on Θ^2 by:

$$\mathcal{K}_T(\theta, \theta') = \langle \phi_T(\theta), \phi_T(\theta') \rangle_{L^2(\lambda_T)}. \tag{4.8}$$

The kernel $\mathcal{K}_T$ measures the colinearity of two features belonging to the continuous dictionary. It does not *a priori* have a simple form. In the following, we approximate this kernel by another kernel easier to handle.

As mentionned in the introduction, we consider in this chapter a setting where the sequence of measures $(\lambda_T, T \geq 1)$ converges towards the Lebesgue measure Leb on Θ. However, since σ_T may drop towards zero, it is often pointless to follow [Butucea et al., 2022a] by taking the pointwise limit kernel of the sequence of kernels $(\mathcal{K}_T,\ T \geq 1)$ as an approximation of the kernel $\mathcal{K}_T$. Indeed, in the next example this pointwise limit kernel is degenerate.

Example 4.2.1 (Degenerate limit kernel). Consider the discrete-time process presented in Section 4.1.2.1 and the Gaussian features (4.6) from Section 4.2.2 scaled by the sequence $(\sigma_T, T \geq 1)$ that tends towards zero when T grows to infinity so that $\lim_{T\to+\infty} \Delta_T/\sigma_T = 0$. In this case, the sequence of measures $(\lambda_T, T \geq 1)$ converges with respect to the vague topology towards the Lebesgue measure and it is easy to check that the pointwise limit of the kernel $\mathcal{K}_T$ is equal to zero almost everywhere.

In what follows, we shall approximate the kernel $\mathcal{K}_T$ by a kernel $\mathcal{K}_T^{\text{prox}}$ of the form:

$$\mathcal{K}_T^{\text{prox}} : (\theta, \theta') \mapsto F(|\theta - \theta'|/\sigma_T), \tag{4.9}$$

where F is a real-valued even function defined on $\mathbb{R}$ with $F(0) = 1$. Since F is even, notice that if it is of class $\mathcal{C}^{2\ell}$ then $\mathcal{K}_T^{\text{prox}}$ is of class $\mathcal{C}^{\ell,\ell}$. The choice of the function F follows from the model given by h, so that $\mathcal{K}_T$ and $\mathcal{K}_T^{\text{prox}}$ are close (see *(iii)* of Assumption 4.2.4). We refer to Sections 4.5 and 4.6 for examples with h given by (4.6) and (4.7). The introduction of the kernel $\mathcal{K}_T^{\text{prox}}$ is significantly different from the approximation developed in [Butucea et al., 2022a].

4.2.3.2 Covariant derivatives of the kernel

Let $\mathcal{K}$ be a symmetric kernel of class $\mathcal{C}^2$ such that the function $g_{\mathcal{K}}$ defined on Θ by:

$$g_{\mathcal{K}}(\theta) = \partial^2_{x,y}\mathcal{K}(\theta, \theta), \tag{4.10}$$

is positive and locally bounded, where ∂_x (respectively ∂_y) denotes the usual derivative with respect to the first (respectively second) variable. Under Assumptions 4.2.1 and 4.2.2, the definitions (4.5) and (4.10) coincide so that $g_T = g_{\mathcal{K}_T}$ on Θ.

Similarly to [Poon et al., 2021], we introduce the covariant derivatives which reduce to elementary expressions since the location parameters are one-dimensional. More precisely following [Butucea et al., 2022a, Section 4], we set for a smooth function f defined on Θ, $\tilde{D}_{0;\mathcal{K}}[f] = f$, $\tilde{D}_{1;\mathcal{K}}[f] = g_{\mathcal{K}}^{-1/2} f'$ and for $i \geq 2$:

$$\tilde{D}_{i;\mathcal{K}}[f] = \tilde{D}_{1;\mathcal{K}}[\tilde{D}_{i-1;\mathcal{K}}[f]].$$

Let us assume that the kernel $\mathcal{K}$ has the form $\mathcal{K}(\theta, \theta') = \langle f(\theta), f(\theta')\rangle_{L^2(\lambda)}$ for some function f of class $\mathcal{C}^3$ and some measure λ on Θ. We then define the covariant derivatives of $\mathcal{K}$ for $i, j \in \{0, \ldots, 3\}$ and $\theta, \theta' \in \Theta$ by:

$$\mathcal{K}^{[i,j]}(\theta, \theta') = \langle \tilde{D}_{i;\mathcal{K}}[f](\theta), \tilde{D}_{j;\mathcal{K}}[f](\theta')\rangle_{L^2(\lambda)}. \tag{4.11}$$

We also define the function $h_{\mathcal{K}}$ on Θ by:

$$h_{\mathcal{K}}(\theta) = \mathcal{K}^{[3,3]}(\theta, \theta).$$

Before stating technical assumptions on the function F, we set:

$$g_\infty = -F''(0). \tag{4.12}$$

For a real valued function f defined on a set A, we write $\|f\|_\infty = \sup_{x\in A} |f(x)|$.

Assumption 4.2.3 (Properties of the function F)**.** *We assume that the function F is of class $\mathcal{C}^6$ and that we have:*

$$g_\infty > 0, \quad L_6 := g_\infty^{-3}|F^{(6)}(0)| < +\infty, \quad \textit{and} \quad L_i := g_\infty^{-i/2} \left\|F^{(i)}\right\|_\infty < +\infty \quad \forall i \in \{0, \cdots, 4\}. \tag{4.13}$$

We give the covariant derivatives of the kernel $\mathcal{K}_T^{\text{prox}}$ according to the definition given in [Butucea et al., 2022a, (27)]. This definition coincides with (4.11) when $\mathcal{K}_T^{\text{prox}}(\theta,\theta') = \langle f(\theta), f(\theta')\rangle_{L^2(\lambda)}$ on Θ^2 for some smooth function f and some measure λ on Θ, see [Butucea et al., 2022a, Lemma 4.3]. We get for any $\theta,\theta' \in \Theta$ and $i,j \in \{0,\cdots,3\}$:

$$\mathcal{K}_T^{\text{prox}[i,j]}(\theta,\theta') = \frac{(-1)^j}{g_\infty^{(i+j)/2}} F^{(i+j)}\left(|\theta-\theta'|/\sigma_T\right). \tag{4.14}$$

We notice that we have for any $\theta \in \Theta$:

$$g_{\mathcal{K}_T^{\text{prox}}}(\theta) = g_\infty/\sigma_T^2. \tag{4.15}$$

4.2.3.3 Measuring the quality of the approximation

In this section, we quantify the proximity of the kernel $\mathcal{K}_T$ and $\mathcal{K}_T^{\text{prox}}$.

Following [Poon et al., 2021], we define the one-dimensional Riemannian metric $\mathfrak{d}_T(\theta,\theta')$ between $\theta,\theta' \in \Theta$ by:

$$\mathfrak{d}_T(\theta,\theta') = |G_T(\theta) - G_T(\theta')|, \tag{4.16}$$

where G_T is a primitive of the function $\sqrt{g_T}$ assumed positive on Θ thanks to Assumption 4.2.2.

Recall that Θ_T, introduced below the model (4.2), is a compact sub-interval of Θ. Since Θ_T is compact, under Assumptions 4.2.2 and 4.2.3, we deduce that the constant C_T below is positive and finite, where:

$$C_T = \max\left(\sup_{\Theta_T}\sqrt{\frac{g_{\mathcal{K}_T^{\text{prox}}}}{g_T}}, \sup_{\Theta_T}\sqrt{\frac{g_T}{g_{\mathcal{K}_T^{\text{prox}}}}}\right). \tag{4.17}$$

Elementary calculations show that the metric $\mathfrak{d}_T$ defined in (4.16) is equivalent, up to a factor σ_T, to the Euclidean metric on Θ_T^2 as for any $\theta,\theta' \in \Theta_T$:

$$\frac{1}{C_T}\sqrt{g_\infty}\,\sigma_T^{-1}\,|\theta-\theta'| \le \mathfrak{d}_T(\theta,\theta') \le C_T\sqrt{g_\infty}\,\sigma_T^{-1}\,|\theta-\theta'|. \tag{4.18}$$

In order to quantify the approximation of $\mathcal{K}_T$ by $\mathcal{K}_T^{\text{prox}}$, we set:

$$\mathcal{V}_T = \max(\mathcal{V}_T^{(1)}, \mathcal{V}_T^{(2)}), \tag{4.19}$$

$$\text{with}\quad \mathcal{V}_T^{(1)} = \max_{i,j\in\{0,1,2\}} \sup_{\Theta_T^2} |\mathcal{K}_T^{[i,j]} - \mathcal{K}_T^{\text{prox}[i,j]}| \quad\text{and}\quad \mathcal{V}_T^{(2)} = \sup_{\Theta_T} |h_{\mathcal{K}_T} - h_{\mathcal{K}_T^{\text{prox}}}|.$$

4.2.4 Boundedness and local concavity on the diagonal of the approximating kernel

Recall the definition of the kernel $\mathcal{K}_T^{\text{prox}}$ given by (4.9) using the even function F. We quantify the boundedness and local concavity on the diagonal of the kernel $\mathcal{K}_T^{\text{prox}}$ using for $r > 0$:

$$\begin{aligned}\varepsilon(r) &= 1 - \sup\left\{|F(r')|;\quad r' \ge r\right\},\\ \nu(r) &= -\sup\left\{F''(r')/g_\infty;\quad r' \in [0,r]\right\}.\end{aligned}$$

We also quantify the colinearity between $s \in \mathbb{N}$ features belonging to the continuous dictionary, by setting for $u > 0$:

$$\delta(u,s) = \inf\Big\{\delta > 0 : \max_{1\le\ell\le s}\sum_{k=1,k\neq\ell}^{s} g_\infty^{-\frac{i}{2}}|F^{(i)}(x_\ell - x_k)| \le u,$$
$$\text{for all } i \in \{0,1,2,3\} \text{ and } (x_1,\cdots,x_s) \in \mathbb{R}^s(\delta)\Big\}, \tag{4.20}$$

where for any subset A of $\mathbb{R}$ or $\mathbb{R}/\mathbb{Z}$ and for any $\delta \geq 0$,

$$A^s(\delta) = \Big\{(\theta_1, \cdots, \theta_s) \in A^s : \, |\theta_\ell - \theta_k| > \delta \text{ for all distinct } k, \ell \in \{1, \ldots, s\}\Big\}. \tag{4.21}$$

with the conventions $\inf \emptyset = +\infty$, and for $s = 0, 1$: $A^0(\delta) = \{0\}$ and $A^1(\delta) = A$.

Following [Butucea et al., 2022a], we define quantities which depend only on the function F and on a real parameter $r > 0$:

$$H_\infty^{(1)}(r) = \frac{1}{2} \wedge L_2 \wedge L_3 \wedge L_4 \wedge L_6 \wedge \frac{\nu(2r)}{10} \wedge \frac{\varepsilon(r/2)}{10},$$
$$H_\infty^{(2)}(r) = \frac{1}{6} \wedge \frac{8\varepsilon(r/2)}{10(5 + 2L_1)} \wedge \frac{8\nu(2r)}{9(2L_2 + 2L_3 + 4)},$$

where the constants L_i are defined in (4.13).

Under Assumption 4.2.4 defined below, we shall build consistent estimators for $\beta^\star$ and $\vartheta^\star$ of the model (4.2).

Assumption 4.2.4. *Let $T \in \mathbb{N}$, $s \in \mathbb{N}$, $r \in \big(0, 1/\sqrt{2\, g_\infty\, L_2}\big)$, $\eta \in (0, 1)$ and a subset $\mathcal{Q} \subset \Theta_T$ of cardinal s.*

(i) ***Regularity of the dictionary φ_T:*** *The dictionary function φ_T satisfies the smoothness conditions of Assumption 4.2.1. The function g_T defined in (4.5), satisfies the positivity condition of Assumption 4.2.2.*

(ii) ***Properties of the function F:*** *Assumption 4.2.3 holds and we have $\varepsilon(r/2) > 0$ and $\nu(2r) > 0$.*

(iii) ***Proximity to the limit setting:*** *The kernel $\mathcal{K}_T$ defined from the dictionary, see (4.8), is sufficiently close to the kernel $\mathcal{K}_T^{prox}$ in the sense that we have:*

$$C_T \leq 2$$

and if $s \geq 1$, we have in addition:

$$\mathcal{V}_T \leq H_\infty^{(1)}(r) \quad \text{and} \quad (s-1)\mathcal{V}_T \leq (1-\eta) H_\infty^{(2)}(r).$$

(iv) ***Separation of the non-linear parameters:*** *If $s \geq 1$, we have:*

$$\delta(\eta H_\infty^{(2)}(r), s) < +\infty \quad \text{and for any } \theta \neq \theta' \in \mathcal{Q}, \quad |\theta - \theta'| > \sigma_T\, \Sigma(\eta, r, s),$$

where,

$$\Sigma(\eta, r, s) = 4 \, \max\Big(r g_\infty^{-1/2}, 2\, \delta(\eta H_\infty^{(2)}(r), s)\Big).$$

Remark 4.2.2 (On the separation). We shall perfom the estimation of $\beta^\star$ and $\vartheta^\star = (\theta_1^\star, \cdots, \theta_s^\star)$ from model (4.2) under the separation condition:

$$|\theta_k^\star - \theta_\ell^\star| \geq \sigma_T\, \Sigma(\eta, r, s), \quad \text{for all } 1 \leq k, \ell \leq s,\ k \neq \ell, \tag{4.22}$$

with $\Sigma(\eta, r, s)$ given in *(iv)* of Assumption 4.2.4. Taking into account the separation condition, the number of admissible features which can be used for the prediction is at most of order $|\Theta_T|/\sigma_T$; this provides a natural upper bound on s. As η is usually fixed, we highlight that the least separation bound tends towards zero when the scaling σ_T goes down to zero.

4.2.5 Prediction error bound

We define the estimators $\hat{\beta}$ and $\hat{\vartheta}$ of $\beta^\star$ and $\vartheta^\star$ as the solution to the following regularized optimization problem with a real tuning parameter $\kappa > 0$ and a bound K on the unknown number s of active features in the observed mixture:

$$(\hat{\beta}, \hat{\vartheta}) \in \underset{\beta \in \mathbb{R}^K, \vartheta \in \Theta_T^K}{\operatorname{argmin}} \quad \frac{1}{2}\|y - \beta \Phi_T(\vartheta)\|^2_{L^2(\lambda_T)} + \kappa \|\beta\|_{\ell_1}, \tag{4.23}$$

where $\|\cdot\|_{\ell_1}$ corresponds to the usual ℓ_1 norm. Since the interval Θ_T on which the optimization of the non-linear parameters is performed is a compact interval and the function Φ_T is continuous, the existence of at least a solution is guaranteed. The bound K on the number s of features in the mixture from model (4.2) allows to formulate an optimization problem. It can be arbitrarily large. In particular, it is not involved in the bounds on estimation and prediction risks given in [Butucea et al., 2022a] with high probability (see Remark 2.4 therein). We stress that the constants in [Butucea et al., 2022a] appearing in those bounds may *a priori* depend on T when the features are scaled by σ_T. We show below that, in fact, those bounds still hold with constants free of T. The results in [Butucea et al., 2022a] as well as the proof of Theorem 4.2.3 below rely on the existence of certificate functions. In [Butucea et al., 2022a], sufficient conditions for the certificate functions to exist are given, see Proposition 7.4 and 7.5 therein. Those conditions require the non-linear parameters in $\mathcal{Q}^\star$ to satisfy the separation condition (4.22). In our framework where the scaling σ_T decreases to zero, it turns out that this separation is in general increasing with s and decreasing with T. However, for some dictionary composed of translated spikes that vanish quickly, it converges to zero when both s and T grow to infinity. We refer to Section 4.5 in this direction.

Recall the definitions of g_∞ and L_2 given by (4.12) and (4.13). The following theorem is a variation of [Butucea et al., 2022a, Theorem 2.1].

Theorem 4.2.3. *Let $T \in \mathbb{N}, s \in \mathbb{N}^*$, $K \in \mathbb{N}^*$, $\eta \in (0,1)$, $r \in \big(0, 1/\sqrt{2\, g_\infty\, L_2}\big)$. Assume we observe the random element y of $L^2(\lambda_T)$ under the regression model (4.2) with unknown parameters $\beta^\star \in (\mathbb{R}^*)^s$ and $\vartheta^\star = (\theta_1^\star, \cdots, \theta_s^\star)$ a vector with distinct entries in Θ_T, a compact interval of Θ, such that Assumption 4.2.4 holds for $\mathcal{Q}^\star = \{\theta_1^\star, \cdots, \theta_s^\star\} \subset \Theta_T$. Assume that the unknown number of active features s is bounded by K. Suppose also that the noise process w_T satisfies Assumption 4.1.1 for a noise level $\overline{\sigma} > 0$ and a decay rate for the noise variance $\Delta_T > 0$.*

Then, there exist finite positive constants $\mathcal{C}_i$, for $i = 0, \ldots, 3$, depending on the function F and on r such that for any $\tau > 1$ and a tuning parameter:

$$\kappa \geq \mathcal{C}_1 \overline{\sigma} \sqrt{\Delta_T \log(\tau)}, \tag{4.24}$$

we have the prediction error bound of the estimators $\hat{\beta}$ and $\hat{\vartheta}$ defined in (4.23) given by:

$$\left\| \hat{\beta} \Phi_T(\hat{\vartheta}) - \beta^\star \Phi_T(\vartheta^\star) \right\|_{L^2(\lambda_T)} \leq \mathcal{C}_0 \sqrt{s}\, \kappa, \tag{4.25}$$

with probability larger than $1 - \mathcal{C}_2 \left(\frac{|\Theta_T|}{\sigma_T\, \tau \sqrt{\log(\tau)}} \vee \frac{1}{\tau} \right)$ where $|\Theta_T|$ is the Euclidean length of Θ_T. Moreover, with the same probability, the difference of the ℓ_1-norms of $\hat{\beta}$ and $\beta^\star$ is bounded by:

$$\left| \|\hat{\beta}\|_{\ell_1} - \|\beta^\star\|_{\ell_1} \right| \leq \mathcal{C}_3\, \kappa\, s. \tag{4.26}$$

Proof. The proof is similar to the proof of [Butucea et al., 2022a, Theorem 2.1] where one replaces the limit kernel noted $\mathcal{K}_\infty$ therein by the approximating kernel $\mathcal{K}_T^{\text{prox}}$ defined in (4.9). The main difference is in checking condition (v) in Theorem 2.1 on the existence of certificate functions. This is done by using Propositions 7.4 and 7.5 therein, and by noticing that the

special form of the approximating kernel $\mathcal{K}_T^{\text{prox}}$ implies that the constants involved do not depend on the scale parameter σ_T. Indeed Equation (4.14) clearly entails that they do not depend on the scale parameter. The details of the proof are left to the interested reader. Details are given in Section 4.7. □

Notice that even if the constants $\mathcal{C}_i$, for $i = 0, \dots, 3$, depend only the function F and on r, Assumption 4.2.4 *(iii)* implies that F is chosen according to the function h. The estimation risks on $\beta^\star$ and $\vartheta^\star$ can be further deduced as in [Butucea et al., 2022a, Equations (9-10)].

The following lemma gives an identifiability result for the considered model. It relies on the construction of certificates from [Butucea et al., 2022a] and is based on ideas developed in [de Castro and Gamboa, 2012] for exact reconstruction of measures, see Lemma 1.1 therein. We recall that by convention $\beta^\star \Phi_T(\vartheta^\star) = 0$ when $s = 0$.

Lemma 4.2.4 (Sufficient conditions for identifiability)**.** *Let $T \in \mathbb{N}$, $r \in \big(0, 1/\sqrt{2\, g_\infty\, L_2}\big)$, $\eta \in (0,1)$. Suppose that Assumption 4.2.4 holds for the set $\mathcal{Q}^\star = \{\theta_1^\star, \cdots, \theta_s^\star\} \subset \Theta_T$ of cardinal $s \in \mathbb{N}$ and for the set $\mathcal{Q}^0 = \{\theta_1^0, \cdots, \theta_{s^0}^0\} \subset \Theta_T$ of cardinal $s^0 \in \mathbb{N}$. Then, for any vectors $\beta^\star \in (\mathbb{R}^*)^s, \beta^0 \in (\mathbb{R}^*)^{s^0}$, we have that, up to the same permutation on the components of $\beta^\star$ and $\vartheta^\star$:*

$$\beta^\star \Phi_T(\vartheta^\star) = \beta^0 \Phi_T(\vartheta^0) \quad \textit{in } L^2(\lambda_T), \qquad \textit{implies that} \qquad s = s^0, \quad \beta^\star = \beta^0 \quad \textit{and} \quad \vartheta^\star = \vartheta^0.$$

Remark 4.2.5. Recall that if $s \geq 1$, then $\beta^\star$ is a s-dimensional vector with non-zero entries. Under the assumptions of Lemma 4.2.4 we have that:

$$\beta^\star \Phi_T(\vartheta^\star) = 0 \quad \text{if and only if} \quad s = 0.$$

Remark 4.2.6. Notice that $\beta^\star \Phi_T(\vartheta^\star) = \beta^0 \Phi_T(\vartheta^0)$ can be re-written as $\tilde{\beta} \Phi_T(\tilde{\vartheta}) = 0$ for some $(\tilde{\beta}, \tilde{\vartheta}) \in \mathbb{R}^{\tilde{s}} \times \Theta_T^{\tilde{s}}$ where the components of $\tilde{\vartheta}$ are the elements of $\mathcal{Q}^\star \cup \mathcal{Q}^0$, $\tilde{s} = \text{Card}(\mathcal{Q}^\star \cup \mathcal{Q}^0)$ and the entries of $\tilde{\beta}$ are up to a sign those of $\beta^\star$ or β^0. In fact, one could show Lemma 4.2.4 by supposing that Assumption 4.2.4 stands for the set $\mathcal{Q}^\star \cup \mathcal{Q}^0$. However, as Assumption 4.2.4 requires pairwise separations between the considered location parameters (see *(iv)* of Assumption 4.2.4), we remark that this condition would be much stronger than requiring that the sets $\mathcal{Q}^\star$ and $\mathcal{Q}^0$ verify Assumption 4.2.4 separately.

Proof of Lemma 4.2.4. First, for $s \geq 1$ and $\vartheta^\star = (\theta_1^\star, \cdots, \theta_s^\star)$ such that Assumption 4.2.4 stands for the set $\mathcal{Q}^\star$, we show that the application $\beta \mapsto \beta \Phi_T(\vartheta^\star)$ defined from $\mathbb{R}^s$ to $L^2(\lambda_T)$ is injective.

We have that $\|\beta \Phi_T(\vartheta^\star)\|_{L^2(\lambda_T)} = \beta \Gamma \beta^\top$, where $\Gamma \in \mathbb{R}^{s \times s}$ is the symmetric matrix defined by $\Gamma_{k,\ell} = \mathcal{K}_T(\theta_k^\star, \theta_\ell^\star)$. Let $\lambda_{\min}$ be the smallest eigenvalue of Γ. Using Gershgorin's theorem and the definition of $\mathcal{V}_T$ given by (4.19), we have that:

$$\lambda_{\min} \geq 1 - \max_{1 \leq \ell \leq s} \sum_{k=1, k \neq \ell}^{s} |\mathcal{K}_T(\theta_\ell^\star, \theta_k^\star)| \geq 1 - \max_{1 \leq \ell \leq s} \sum_{k=1, k \neq \ell}^{s} \left| F\left(\frac{|\theta_\ell^\star - \theta_k^\star|}{\sigma_T} \right) \right| - (s-1)\mathcal{V}_T.$$

The separation condition from Point *(iv)* of Assumption 4.2.4 implies that for all $k, \ell \in \{1, \cdots, s\}$ such that $k \neq \ell$ we have $|\theta_k^\star - \theta_\ell^\star| \geq \sigma_T \Sigma(\eta, r, s) \geq 8\, \sigma_T\, \delta(\eta H_\infty^{(2)}(r), s)$. Recall the definition of $\delta(u, s)$ given by (4.20). We deduce that:

$$\max_{1 \leq \ell \leq s} \sum_{k=1, k \neq \ell}^{s} \left| F\left(\frac{|\theta_\ell^\star - \theta_k^\star|}{\sigma_T} \right) \right| \leq \eta H_\infty^{(2)}(r).$$

By Point *(iii)* of Assumption 4.2.4, we have $(s-1)\mathcal{V}_T \leq (1-\eta) H_\infty^{(2)}(r)$ and $H_\infty^{(2)}(r) \leq 1/6$. Thus, we get:

$$\lambda_{\min} \geq 5/6. \tag{4.27}$$

Hence, the symmetric matrix Γ is positive-definite. This proves that the application $\beta \mapsto \beta\Phi_T(\vartheta^\star)$ is injective from $\mathbb{R}^s$ to $L^2(\lambda_T)$. By symmetry, we obtain for $s^0 \geq 1$ that the application $\beta \mapsto \beta\Phi_T(\vartheta^0)$ is injective from $\mathbb{R}^{s^0}$ to $L^2(\lambda_T)$.

If $s = 0$, we have $\beta^\star\Phi_T(\vartheta^\star) = 0$. For $s^0 \geq 1$, we have $\beta^0 \in (\mathbb{R}^*)^{s^0}$ and since $\beta \mapsto \beta\Phi_T(\vartheta^0)$ is injective, we deduce that $\beta^0\Phi_T(\vartheta^0) \neq 0$. Thus, $s = 0$ and $\beta^\star\Phi_T(\vartheta^\star) = \beta^0\Phi_T(\vartheta^0)$ implies that $s^0 = 0$. By symmetry, $s^0 = 0$ and $\beta^\star\Phi_T(\vartheta^\star) = \beta^0\Phi_T(\vartheta^0)$ implies also that $s = 0$.

Assume from now on that $s, s^0 \in \mathbb{N}^*$ and that $\beta^\star\Phi_T(\vartheta^\star) = \beta^0\Phi_T(\vartheta^0)$. Let us consider the application $v : \mathcal{Q}^\star \mapsto \{-1, 1\}$ defined by: $v(\theta_k^\star) = \mathrm{sign}(\beta_k^\star)$ for any $k \in \{1, \cdots, s\}$. According to Lemma 4.4.2, there exists $p^\star \in L^2(\lambda_T)$ such that:

$$\|\beta^\star\|_{\ell_1} = \sum_{k=1}^{s} \beta_k^\star \langle \phi_T(\theta_k^\star), p^\star \rangle_{L^2(\lambda_T)} = \langle \beta^\star\Phi_T(\vartheta^\star), p^\star \rangle_{L^2(\lambda_T)} .$$

Using the fact that $\beta^\star\Phi_T(\vartheta^\star) = \beta^0\Phi_T(\vartheta^0)$ and Properties (i) and (ii) of $p^\star$ in Lemma 4.4.2, we get:

$$\|\beta^\star\|_{\ell_1} = \sum_{k=1}^{s^0} \beta_k^0 \left\langle \phi_T(\theta_k^0), p^\star \right\rangle_{L^2(\lambda_T)} \leq \left\|\beta^0\right\|_{\ell_1}. \tag{4.28}$$

The role of $(\beta^\star, \vartheta^\star)$ and (β^0, ϑ^0) being symmetric, we also get $\left\|\beta^0\right\|_{\ell_1} \leq \|\beta^\star\|_{\ell_1}$. Hence, we have $\left\|\beta^0\right\|_{\ell_1} = \|\beta^\star\|_{\ell_1}$ and $\mathrm{sign}(\beta_k^0) = \langle \phi_T(\theta_k^0), p^\star \rangle_{L^2(\lambda_T)}$ for $k \in \{1, \cdots, s^0\}$. Using Properties (i) and (ii) of $p^\star$ in Lemma 4.4.2, we remark that for any $\theta \notin \mathcal{Q}^\star$

$$\left|\langle \phi_T(\theta), p^\star \rangle_{L^2(\lambda_T)}\right| < 1.$$

Thus, we deduce from (4.28) that $\mathcal{Q}^0 \subseteq \mathcal{Q}^\star$ and by symmetry $\mathcal{Q}^0 = \mathcal{Q}^\star$. Hence, we obtain $\vartheta^\star = \vartheta^0$ (up to a permutation on the components of $\vartheta^\star$) and $s = s^0$. Then use the injectivity of the function $\beta \mapsto \beta\Phi_T(\vartheta^\star)$ to get that $\beta^\star = \beta^0$ (up to the same permutation). This finishes the proof of the Lemma. □

4.3 Goodness-of-fit for the mixture model

In this section, we build a test procedure to decide if the observation y derives from a given mixture of translated features. We build a test Ψ, *i.e.* a measurable function of the observation y taking value in $\{0, 1\}$, in order to distinguish a null hypothesis H_0 against an alternative $H_1(\rho)$ depending on a nonnegative separation parameter ρ. We recall that the maximal type I and II error probabilities are $\sup_{(\beta^\star,\vartheta^\star)\in H_0} \mathbb{E}_{(\beta^\star,\vartheta^\star)}[\Psi]$ and $\sup_{(\beta^\star,\vartheta^\star)\in H_1(\rho)} \mathbb{E}_{(\beta^\star,\vartheta^\star)}[1 - \Psi]$, respectively, where Ψ is a function of y which is equal to $\beta^\star\Phi_T(\vartheta^\star) + w_T$ under $\mathbb{E}_{(\beta^\star,\vartheta^\star)}$. The maximal testing risk is the sum of the former quantities, that is:

$$R_\rho(\Psi) = \sup_{(\beta^\star,\vartheta^\star)\in H_0} \mathbb{E}_{(\beta^\star,\vartheta^\star)}[\Psi] + \sup_{(\beta^\star,\vartheta^\star)\in H_1(\rho)} \mathbb{E}_{(\beta^\star,\vartheta^\star)}[1 - \Psi],$$

and the minimax testing risk is:

$$R_\rho^\star = \inf_{\Psi} R_\rho(\Psi), \tag{4.29}$$

where the infinimum is taken over all the measurable functions from $L^2(\lambda_T)$ to $\{0, 1\}$. The minimax separation rate of the test problem is defined for any $\alpha \in (0, 1)$ as:

$$\rho^\star(\alpha) = \inf\{\rho > 0 : R_\rho^\star \leq \alpha\}. \tag{4.30}$$

4.3.1 Test problem

Let $s^0 \in \mathbb{N}$ and consider the set $\Theta_T^{s^0}(\delta^0) \subset \Theta_T^{s^0}$ of vectors whose components are pairwise separated by a distance $\delta^0 \geq 0$ (recall the definition (4.21)). Consider the vectors $\beta^0 \in (\mathbb{R}^*)^{s^0}$ and $\vartheta^0 = (\theta_1^0, \cdots, \theta_{s^0}^0) \in \Theta_T^{s^0}(\delta^0)$. By convention, we have for $s^0 = 0$ that $\beta^0 = 0$, $\vartheta^0 = 0$ and $\beta^0 \Phi_T(\vartheta^0) = 0$.

We build a test procedure based on the observation y to decide, for some $\delta^\star \geq 0$, whether:

$$\begin{cases} H_0 : & (\beta^\star, \vartheta^\star) \in (\mathbb{R}^*)^s \times \Theta_T^s(\delta^\star) \quad \text{such that} \quad \beta^\star \Phi_T(\vartheta^\star) = \beta^0 \Phi_T(\vartheta^0), \\ H_1(\rho) : & (\beta^\star, \vartheta^\star) \in (\mathbb{R}^*)^s \times \Theta_T^s(\delta^\star) \quad \text{such that} \quad \left\| \beta^\star \Phi_T(\vartheta^\star) - \beta^0 \Phi_T(\vartheta^0) \right\|_{L^2(\lambda_T)} \geq \rho, \end{cases} \tag{4.31}$$

where ρ is a nonnegative separation parameter. When Assumption 4.2.4 holds for the sets $\mathcal{Q}^\star = \{\theta_1^\star, \cdots, \theta_s^\star\}$ and $\mathcal{Q}^0 = \{\theta_1^0, \cdots, \theta_{s^0}^0\}$, by Lemma 4.2.4, the null hypothesis implies that $(\beta^\star, \vartheta^\star) = (\beta^0, \vartheta^0)$ (up to the same permutation on the components of $\beta^\star$ and $\vartheta^\star$). We remark that the separation condition from Point *(iv)* of Assumption 4.2.4 required between the elements of $\mathcal{Q}^\star$ (resp. $\mathcal{Q}^0$) is automatically satisfied when $\delta^\star \geq \sigma_T \, \Sigma(\eta, r, s)$ (resp. $\delta^0 \geq \sigma_T \, \Sigma(\eta, r, s^0)$).

We shall denote the distribution under the null hypothesis as associated to the parameters (β^0, ϑ^0) and see that the maximal type I error probability writes in this case $\mathbb{E}_{(\beta^0, \vartheta^0)}[\Psi]$ for $\mathbb{E}_{(\beta^\star, \vartheta^\star)}[\Psi]$. Furthermore, when $s^0 = 0$, under Assumption 4.2.4 for the set $\mathcal{Q}^\star$, Lemma 4.2.4 implies that the null hypothesis reduces to $H_0 : s = 0$.

4.3.2 Main results

We consider the test procedure $\Psi_{\mathcal{T}}(t)$ associated to a real valued statistic $\mathcal{T}$ (measurable function of the observation y) and a threshold $t > 0$ (defining a critical region) given by:

$$\Psi_{\mathcal{T}}(t) = \mathbf{1}_{\{|\mathcal{T}| > t\}}. \tag{4.32}$$

We recall that for a test Ψ, we accept H_0 when $\Psi = 0$ and reject it when $\Psi = 1$.

Let $s^0 \in \mathbb{N}$ and consider known linear coefficients and location parameters $\beta^0 \in (\mathbb{R}^*)^{s^0}$ and $\vartheta^0 = (\theta_1^0, \cdots, \theta_{s^0}^0) \in \Theta_T^{s^0}$, respectively. We define two statistics $\mathcal{T}_1$ and $\mathcal{T}_2$ by:

$$\mathcal{T}_1 = \left\| y - \beta^0 \Phi_T(\vartheta^0) \right\|_{L^2(\lambda_T)}^2 - \mathbb{E}\left[\|w_T\|_{L^2(\lambda_T)}^2 \right] \quad \text{and} \quad \mathcal{T}_2 = \left\| \hat{\beta} \Phi_T(\hat{\vartheta}) - \beta^0 \Phi_T(\vartheta^0) \right\|_{L^2(\lambda_T)}^2, \tag{4.33}$$

where $\hat{\beta}$ and $\hat{\vartheta}$ denote the estimators obtained from (4.23) for a given value of the tuning parameter κ and a bound K on the unknown number $s \in \mathbb{N}$ of active features in the observed signal.

Recall the definition (4.3) of Ξ_T, the variance of the squared $L^2(\lambda_T)$-norm of the noise w_T. The following theorem gives an upper bound of the maximal testing risk associated to the test $\Psi_{\mathcal{T}_1}(t)$ for some positive threshold t and positive separation ρ.

Theorem 4.3.1. *Let $T \in \mathbb{N}$ and $s^0 \in \mathbb{N}$. Let $\delta^\star \geq 0$ and $\delta^0 \geq 0$. Assume that we observe the random element y of $L^2(\lambda_T)$ under the regression model (4.2) with unknown parameters $s \in \mathbb{N}$, $\beta^\star \in (\mathbb{R}^*)^s$ and $\vartheta^\star \in \Theta_T^s(\delta^\star)$. Let $\beta^0 \in (\mathbb{R}^*)^{s^0}$ and $\vartheta^0 \in \Theta_T^{s^0}(\delta^0)$. Suppose that Assumption 4.2.1 on the smoothness of the features holds. Suppose that Assumption 4.1.1 holds for a noise level $\overline{\sigma} > 0$ and a decay rate for the noise variance $\Delta_T > 0$.*

Then, the test $\Psi_{\mathcal{T}_1}$ in (4.32) using $\mathcal{T}_1$ in (4.33) satisfies:

$$R_\rho\left(\Psi_{\mathcal{T}_1}(t)\right) \leq \frac{\Xi_T}{t^2} + \frac{4\,\Xi_T}{(\rho^2 - t)^2} + \mathrm{e}^{-(\rho^2 - t)^2 / (32 \overline{\sigma}^2 \Delta_T \rho^2)}, \tag{4.34}$$

for any threshold t and any separation ρ such that $\rho^2 > t > 0$.

Proof. We give a bound of the type I error probability. Using that under H_0 we have $y = \beta^0 \Phi_T(\vartheta^0) + w_T$, we get:

$$\mathbb{E}_{(\beta^0,\,\vartheta^0)}[\Psi_{\mathcal{T}_1}(t)] = \mathbb{P}\left(\Big|\ \|w_T\|^2_{L^2(\lambda_T)} - \mathbb{E}\left[\|w_T\|^2_{L^2(\lambda_T)}\right]\Big| > t\right).$$

Using Chebyshev's inequality, we obtain:

$$\mathbb{E}_{(\beta^0,\,\vartheta^0)}[\Psi_{\mathcal{T}_1}(t)] \leq \frac{\Xi_T}{t^2}. \tag{4.35}$$

We now give a bound of the type II error probability. We set:

$$R = \left\|\beta^0 \Phi_T(\vartheta^0) - \beta^\star \Phi_T(\vartheta^\star)\right\|_{L^2(\lambda_T)},$$

where $(\beta^\star, \vartheta^\star) \in (\mathbb{R}^*)^s \times \Theta^s_T(\delta^\star)$. Using the decomposition of y from the model (4.2) and the triangle inequality, we have:

$$|\mathcal{T}_1| \geq R^2 - \left|\|w_T\|^2_{L^2(\lambda_T)} - \mathbb{E}[\|w_T\|^2_{L^2(\lambda_T)}]\right| - 2\left|\left\langle \beta^0 \Phi_T(\vartheta^0) - \beta^\star \Phi_T(\vartheta^\star), w_T \right\rangle_{L^2(\lambda_T)}\right|.$$

Notice that by Assumption 4.1.1, the random variable $\langle \beta^0 \Phi_T(\vartheta^0) - \beta^\star \Phi_T(\vartheta^\star), w_T \rangle_{L^2(\lambda_T)}$ is Gaussian with zero mean and variance bounded by $\overline{\sigma}^2\, \Delta_T\, R^2$. Hence, using that under $H_1(\rho)$ we have $R \geq \rho$, we obtain:

$$\begin{aligned}\mathbb{E}_{(\beta^\star,\,\vartheta^\star)}[1 - \Psi_{\mathcal{T}_1}(t)] \leq & \mathbb{P}\left((\rho^2 - t)/2 \leq \left|\|w_T\|^2_{L^2(\lambda_T)} - \mathbb{E}[\|w_T\|^2_{L^2(\lambda_T)}]\right|\right) \\ & + \mathbb{P}\left((R^2 - t)/2 \leq 2\overline{\sigma}\sqrt{\Delta_T}\, R\, |G|\right),\end{aligned} \tag{4.36}$$

where G is a standard Gaussian random variable. On the one hand, for $t < \rho^2$, using Chebyshev's inequality we get:

$$\mathbb{P}\left((\rho^2 - t)/2 \leq \left|\|w_T\|^2_{L^2(\lambda_T)} - \mathbb{E}[\|w_T\|^2_{L^2(\lambda_T)}]\right|\right) \leq \frac{4\,\Xi_T}{(\rho^2 - t)^2}. \tag{4.37}$$

On the other hand, we have:

$$\mathbb{P}\left((R^2 - t)/2 \leq 2\overline{\sigma}\sqrt{\Delta_T}\, R\, |G|\right) \leq \mathbb{P}\left(\frac{\rho^2 - t}{4\overline{\sigma}\sqrt{\Delta_T}\rho} \leq |G|\right) \leq \mathrm{e}^{-(\rho^2 - t)^2/(32\overline{\sigma}^2 \Delta_T \rho^2)}. \tag{4.38}$$

where we used that $\rho \leq R$ and the tail bound (see [Abramowitz and Stegun, 1992, Formula 7.1.13]):

$$\frac{1}{\sqrt{2\pi}} \int_u^{+\infty} \mathrm{e}^{-t^2/2}\, \mathrm{d}t \leq \frac{1}{2}\, \mathrm{e}^{-u^2/2}, \quad \text{for } u > 0. \tag{4.39}$$

By combining (4.36) with (4.37) and (4.38), we get the following bound on the type II error probability:

$$\mathbb{E}_{(\beta^\star,\,\vartheta^\star)}[1 - \Psi_{\mathcal{T}_1}(t)] \leq \frac{4\,\Xi_T}{(\rho^2 - t)^2} + \mathrm{e}^{-(\rho^2 - t)^2/(32\overline{\sigma}^2 \Delta_T \rho^2)}. \tag{4.40}$$

Then, by putting together (4.35) and (4.40), we obtain (4.34). □

We deduce from Theorem 4.3.1 upper bounds on the minimax separation $\rho^\star$ defined in (4.30) for the goodness-of-fit test problem (4.31).

Corollary 4.3.2. *Under the framework and the assumptions of Theorem 4.3.1, the minimax separation rate for the test problem (4.31) verifies for any $\alpha \in (0, 1)$:*

$$\rho^\star(\alpha) \leq \rho^{(1)}(\alpha) \quad \textit{with} \quad \rho^{(1)}(\alpha) := \max\left(\left(\frac{40\Xi_T}{\alpha}\right)^{1/4}, 8\,\overline{\sigma}\sqrt{2\Delta_T \log\left(\frac{2}{\alpha}\right)}\right). \tag{4.41}$$

Proof of Corollary 4.3.2. This result is a direct consequence of Theorem 4.3.1 by taking the threshold t of the test therein equal to $\rho^2/2$. Then, we have that for $\rho > 0$:

$$R_\rho^\star \leq R_\rho\left(\Psi_{\mathcal{T}_1}(\rho^2/2)\right) \leq \frac{4\Xi_T}{\rho^4} + \frac{16\,\Xi_T}{\rho^4} + \mathrm{e}^{-\rho^2/(128\overline{\sigma}^2\Delta_T)} = \frac{20\,\Xi_T}{\rho^4} + \mathrm{e}^{-\rho^2/(128\overline{\sigma}^2\Delta_T)}\,.$$

We deduce that $R_\rho^\star \leq \alpha$ for any $\alpha \in (0,1)$ whenever the separation ρ satisfies:

$$\rho \geq \left(\frac{40\Xi_T}{\alpha}\right)^{\frac{1}{4}} \vee \overline{\sigma}\sqrt{128\,\Delta_T \log\left(\frac{2}{\alpha}\right)}.$$

This implies (4.41). □

In the following theorem, we give a bound of the maximal testing risk associated to the test $\Psi_{\mathcal{T}_2}(t)$ using $\mathcal{T}_2$ in (4.33) for solving the test problem (4.31). The statistic $\mathcal{T}_2$ is defined using estimators of the model parameters $(\beta^\star, \vartheta^\star)$. In view of recovering the latter, we assume that the minimal distance $\delta^\star$ (resp. δ^0) is large enough so that Point *(iv)* of Assumption 4.2.4 is satisfied for the components of $\vartheta^\star$ (resp. ϑ^0).

Recall the definitions of g_∞ and L_2 given by (4.12) and (4.13), that $|\Theta_T|$ denotes the Euclidean length of the compact set Θ_T and Σ defined in *(iv)* of Assumption 4.2.4.

Theorem 4.3.3. *Let $T \in \mathbb{N}$, $s^0 \in \mathbb{N}$ and choose $K \in \mathbb{N}$ such that $s_0 \leq K$. Let also $\eta \in (0,1)$ and $r \in (0, 1/\sqrt{2\,g_\infty\,L_2})$. Let $\delta^\star \geq \sigma_T\,\Sigma(\eta, r, s)$ and $\delta^0 \geq \sigma_T\,\Sigma(\eta, r, s^0)$. Assume we observe the random element y of $L^2(\lambda_T)$ under the regression model (4.2) with unknown parameters $s \in \mathbb{N}$ such that $s \leq K$, $\beta^\star \in (\mathbb{R}^*)^s$ and $\vartheta^\star = (\theta_1^\star, \cdots, \theta_s^\star) \in \Theta_T^s(\delta^\star)$. Let $\beta^0 \in (\mathbb{R}^*)^{s^0}$ and $\vartheta^0 = (\theta_1^0, \cdots, \theta_{s^0}^0) \in \Theta_T^{s^0}(\delta^0)$. Suppose that Assumption 4.2.4 holds for the sets $\mathcal{Q}^\star = \{\theta_1^\star, \cdots, \theta_s^\star\} \subset \Theta_T$ of cardinal s and $\mathcal{Q}^0 = \{\theta_1^0, \cdots, \theta_{s^0}^0\} \subset \Theta_T$ of cardinal s^0. Suppose also that the noise process w_T satisfies Assumption 4.1.1 for a noise level $\overline{\sigma} > 0$ and a decay rate for the noise variance $\Delta_T > 0$.*

Then, there exist finite positive constants $\mathcal{C}_0$, $\mathcal{C}_1$, $\mathcal{C}_2$, depending on r and on the function F, such that for the tuning parameter κ:

$$\kappa \geq \mathcal{C}_1\overline{\sigma}\sqrt{\Delta_T \log(\tau)}, \quad \text{for some } \tau > 1, \tag{4.42}$$

the test $\Psi_{\mathcal{T}_2}$ using $\mathcal{T}_2$ in (4.33) satisfies:

$$R_\rho\left(\Psi_{\mathcal{T}_2}(t)\right) \leq 2\,\mathcal{C}_2\left(\frac{|\Theta_T|}{\sigma_T\,\tau\sqrt{\log(\tau)}} \vee \frac{1}{\tau}\right),$$

for any threshold t and any separation ρ satisfying:

$$0 < t, \quad \mathcal{C}_0\sqrt{s^0}\,\kappa \leq \sqrt{t} < \rho \quad \text{and} \quad \sqrt{t} + \mathcal{C}_0\sqrt{s}\,\kappa \leq \rho. \tag{4.43}$$

Remark 4.3.4 (On the bound K). The bound K on s is assumed to be known. It is needed to formulate the optimization problem (4.23) whose solutions are the estimators of $\beta^\star$ and $\vartheta^\star$. However, we stress that the constants $\mathcal{C}_0$, $\mathcal{C}_1$, $\mathcal{C}_2$ and the bound on the maximal testing risk do not depend on K. Thus, K can be taken arbitrarily large.

Proof of Theorem 4.3.3. **Case $s > 0$.** Let $(\beta^\star, \vartheta^\star) \in (\mathbb{R}^*)^s \times \Theta_T^s(\delta^\star)$. We consider the estimators $(\hat{\beta}, \hat{\vartheta})$ defined in (4.23). Notice that the hypotheses of Theorem 4.2.3 are in force. We use the constants $\mathcal{C}_0$, $\mathcal{C}_1$, $\mathcal{C}_2$ defined therein. Under H_0, we have $s = s^0$. Thus, for $\sqrt{t} \geq \mathcal{C}_0\,\sqrt{s}\,\kappa$, we get the following bound on the type I error probability:

$$\mathbb{E}_{(\beta^0,\,\vartheta^0)}[\Psi_{\mathcal{T}_2}(t)] \leq \mathbb{P}\left(\left\|\hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star\Phi_T(\vartheta^\star)\right\|_{L^2(\lambda_T)} > \mathcal{C}_0\,\sqrt{s}\,\kappa\right) \leq \mathcal{C}_2\left(\frac{|\Theta_T|}{\sigma_T\,\tau\sqrt{\log(\tau)}} \vee \frac{1}{\tau}\right), \tag{4.44}$$

where we used that $\beta^0\Phi_T(\vartheta^0) = \beta^\star\Phi_T(\vartheta^\star)$ and that $\sqrt{t} \geq \mathcal{C}_0\sqrt{s}\,\kappa$ for the first inequality and Theorem 4.2.3 for the second.

We now bound the type II error probability. Under the hypothesis $H_1(\rho)$, since we have $\|\beta^\star\Phi_T(\vartheta^\star) - \beta^0\Phi_T(\vartheta^0)\|_{L^2(\lambda_T)} \geq \rho$, we obtain that:

$$\mathbb{E}_{(\beta^\star,\,\vartheta^\star)}[1-\Psi_{\mathcal{T}_2}(t)] \leq \mathbb{P}\left(\rho - \sqrt{t} \leq \left\|\hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^\star\Phi_T(\vartheta^\star)\right\|_{L^2(\lambda_T)}\right) \leq \mathcal{C}_2\left(\frac{|\Theta_T|}{\sigma_T\,\tau\sqrt{\log(\tau)}} \vee \frac{1}{\tau}\right), \tag{4.45}$$

where we used the triangle inequality for the first inequality and Theorem 4.2.3 as well as $\rho - \sqrt{t} \geq \mathcal{C}_0\sqrt{s}\,\kappa$ for the second.

Case $s = 0$. Since $s = 0$, we have $y = w_T$ according to (4.2). Let us first bound the type I error probability $\mathbb{E}_{(\beta^0,\,\vartheta^0)}[\Psi_{\mathcal{T}_2}(t)]$. Assume that the hypothesis H_0 holds so that $s = s^0 = 0$. By definition we have:

$$\mathbb{E}_{(\beta^0,\,\vartheta^0)}[\Psi_{\mathcal{T}_2}(t)] = \mathbb{P}\left(\left\|\hat{\beta}\Phi_T(\hat{\vartheta})\right\|^2_{L^2(\lambda_T)} > t\right).$$

We get from the definition of the estimators $\hat{\beta}$ and $\hat{\vartheta}$ from (4.23) that:

$$\frac{1}{2}\left\|w_T - \hat{\beta}\Phi_T(\hat{\vartheta})\right\|^2_{L^2(\lambda_T)} + \kappa\left\|\hat{\beta}\right\|_{\ell_1} \leq \frac{1}{2}\|w_T\|^2_{L^2(\lambda_T)}.$$

By rearranging some terms in the equation above, we get:

$$\frac{1}{2}\left\|\hat{\beta}\Phi_T(\hat{\vartheta})\right\|^2_{L^2(\lambda_T)} \leq \left\langle\hat{\beta}\Phi_T(\hat{\vartheta}), w_T\right\rangle_{L^2(\lambda_T)} - \kappa\left\|\hat{\beta}\right\|_{\ell_1} \leq \left\|\hat{\beta}\right\|_{\ell_1}\left(\sup_{\Theta_T}|\,\langle\phi_T(\theta), w_T\rangle_{L^2(\lambda_T)}\,| - \kappa\right). \tag{4.46}$$

Let us define the event:

$$\mathcal{A} = \{\sup_{\theta\in\Theta_T}|\,\langle\phi_T(\theta), w_T\rangle_{L^2(\lambda_T)}\,| < \kappa\}.$$

We deduce from (4.46) that on the event $\mathcal{A}$ we have $\left\|\hat{\beta}\Phi_T(\hat{\vartheta})\right\|_{L^2(\lambda_T)} = 0$. Therefore we get:

$$\mathbb{E}_{(\beta^0,\,\vartheta^0)}[\Psi_{\mathcal{T}_2}(t)] \leq \mathbb{P}\left(\left\|\hat{\beta}\Phi_T(\hat{\vartheta})\right\|_{L^2(\lambda_T)} > 0\right) \leq \mathbb{P}(\mathcal{A}^c). \tag{4.47}$$

We shall bound later $\mathbb{P}(\mathcal{A}^c)$, see (4.49).

We now consider the type II error probability. We asume H_1, that is $\|\beta^0\Phi_T(\vartheta^0)\|_{L^2(\lambda_T)} \geq \rho$. We obtain:

$$\begin{aligned}\mathbb{E}_{(\beta^\star,\,\vartheta^\star)}[1-\Psi_{\mathcal{T}_2}(t)] &= \mathbb{P}\left(\left\|\hat{\beta}\Phi_T(\hat{\vartheta}) - \beta^0\Phi_T(\vartheta^0)\right\|_{L^2(\lambda_T)} \leq \sqrt{t}\right)\\ &\leq \mathbb{P}\left(\rho - \sqrt{t} \leq \left\|\hat{\beta}\Phi_T(\hat{\vartheta}))\right\|_{L^2(\lambda_T)}\right) \leq \mathbb{P}(\mathcal{A}^c).\end{aligned} \tag{4.48}$$

where we used the definition of $\mathcal{T}_2$ and the triangle inequality for the first inequality, the second inequality of (4.47) as well as $\rho - \sqrt{t} > 0$ for the second.

We shall apply [Butucea et al., 2022a, Lemma A.1] to bound $\mathbb{P}(\mathcal{A}^c)$. It amounts to controling the supremum of the Gaussian process $\theta \mapsto \langle\phi_T(\theta), w_T\rangle_{L^2(\lambda_T)}$. Recall that Assumptions 4.2.1 and 4.2.2 hold. The function ϕ_T is of class $\mathcal{C}^1$ from the interval Θ_T to $L^2(\lambda_T)$, with Θ_T a sub-interval of Θ. We have also, with $\phi_T^{[1]} = \tilde{D}_{1;\mathcal{K}_T}[\phi_T]$, that:

$$\|\phi_T(\theta)\|_{L^2(\lambda_T)} = 1 \quad \text{and} \quad \left\|\phi_T^{[1]}(\theta)\right\|^2_{L^2(\lambda_T)} = \mathcal{K}_T^{[1,1]}(\theta,\theta) = 1.$$

Since Assumption 4.1.1 on the noise w_T holds, the hypotheses of [Butucea et al., 2022a, Lemma A.1] hold and we deduce from [Butucea et al., 2022a, Lemma A.1] (with $C_1 = C_2 = 1$ therein) that:

$$\mathbb{P}(\mathcal{A}^c) = \mathbb{P}\left(\sup_{\theta\in\Theta_T} |\langle \phi_T(\theta), w_T\rangle_{L^2(\lambda_T)}| \geq \kappa\right) \leq 3\cdot\left(\frac{2\overline{\sigma}\sqrt{g_\infty}|\Theta_T|\sqrt{\Delta_T}}{\sigma_T\kappa} \vee 1\right) \mathrm{e}^{-\kappa^2/(4\overline{\sigma}^2\Delta_T)},$$

where the diameter $|\Theta_T|_{\mathfrak{d}_T}$ of the set Θ_T with respect to the metric $\mathfrak{d}_T$ is bounded by $2\sqrt{g_\infty}|\Theta_T|/\sigma_T$ using (4.18) and the fact that $C_T \leq 2$. By taking $\kappa \geq 2\overline{\sigma}\sqrt{\Delta_T \log(\tau)}$, we get:

$$\mathbb{P}(\mathcal{A}^c) = \mathbb{P}\left(\sup_{\theta\in\Theta_T} |\langle \phi_T(\theta), w_T\rangle_{L^2(\lambda_T)}| \geq \kappa\right) \leq 3\cdot\left(\frac{\sqrt{g_\infty}|\Theta_T|}{\sigma_T\tau\sqrt{\log(\tau)}} \vee \frac{1}{\tau}\right). \tag{4.49}$$

Notice that the constant $\mathcal{C}_2$ from Theorem 4.2.3 is equal to $2\sqrt{g_\infty}\mathcal{C}_2'$ where $\mathcal{C}_2'$ is given by [Butucea et al., 2022a, $\mathcal{C}_2$ from Eq. (84) therein] and is greater than 3. The constant $\mathcal{C}_2$ depends only on r and the function F. Finally, by putting together (4.44), (4.45), (4.47) and (4.48), we obtain for $\kappa \geq \mathcal{C}_1\overline{\sigma}\sqrt{\Delta_T\log(\tau)}$ (where the constant $\mathcal{C}_1$ is defined in [Butucea et al., 2022a, Proof of Theorem 2.1 (p.32)] and is superior to 4) the bound on the maximal testing risk from Theorem 4.3.3. This finishes the proof. □

In the next Corollary, we obtain an additionnal upper bound on the minimax separation rate.

Corollary 4.3.5. *Under the framework and the assumptions of Theorem 4.3.3 and provided that $|\Theta_T|/\sigma_T \geq 1$, there exist finite positive constants c and C, depending on r and the function F, such that the minimax separation rate for the test problem (4.31) verifies for any $\alpha \in (0,1)$:*

$$\rho^\star(\alpha) \leq \rho^{(2)}(\alpha), \quad \rho^{(2)}(\alpha) := C\,\overline{\sigma}\sqrt{(s\vee s^0\vee 1)\Delta_T \log\left(\frac{c\,|\Theta_T|}{\alpha\,\sigma_T}\right)}. \tag{4.50}$$

Remark 4.3.6 (On the condition $|\Theta_T|/\sigma_T \geq 1$). We recall that the set Θ_T is a compact subset of Θ. In the case where Θ is the torus $\mathbb{R}/\mathbb{Z}$, $\Theta_T = \Theta$ and the scale parameter σ_T tends towards 0 when T grows to infinity, the condition $|\Theta_T|/\sigma_T \geq 1$ is satisfied for T large enough. This condition also holds for T large enough in the Gaussian spikes deconvolution example, with the particular choices for Θ_T and σ_T from Section 4.5, where $\Theta = \mathbb{R}$, $\lim_{T\to+\infty}\Theta_T = \Theta$ and $\lim_{T\to+\infty}\sigma_T = 0$.

Proof of Corollary 4.3.5. Notice that all the assumptions of Theorem 4.3.3 are in force. The result is a direct consequence of Theorem 4.3.3. We fix the tuning parameter $\kappa = \mathcal{C}_1\overline{\sigma}\sqrt{\Delta_T\log(\tau)}$ by taking the equality in (4.42). Then, for

$$\rho \geq \mathcal{C}_0\sqrt{s\vee 1}\,\kappa + \sqrt{t} \quad \text{and} \quad t = \mathcal{C}_0^2\,(s^0\vee 1)\,\kappa^2, \tag{4.51}$$

we have (4.43) (in particular $0 < t < \rho$) and by Theorem 4.3.3 for $\tau > 1$:

$$R_\rho^\star \leq R_\rho\left(\Psi_{\mathcal{T}_2}(t)\right) \leq 2\mathcal{C}_2\left(\frac{|\Theta_T|}{\sigma_T\,\tau\sqrt{\log(\tau)}} \vee \frac{1}{\tau}\right),$$

where the finite positive constants $\mathcal{C}_0$, $\mathcal{C}_1$, $\mathcal{C}_2$, from Theorem 4.3.3 depend on r and F.

Then, taking $\tau = c|\Theta_T|/(\alpha\sigma_T)$ with $c = (2\mathcal{C}_2)\vee\mathrm{e}$ and using that by assumption $|\Theta_T|/\sigma_T \geq 1$, we get for $\rho \geq \sqrt{2}\mathcal{C}_0\mathcal{C}_1\overline{\sigma}\sqrt{(s+s^0)\vee 2}\sqrt{\Delta_T\log(c|\Theta_T|/(\alpha\sigma_T))}$ and $\alpha\in(0,1)$ that $R_\rho^\star \leq \alpha$. We readily deduce (4.50) with $C = 2\mathcal{C}_0\mathcal{C}_1$. □

Remark 4.3.7 (Combining the upper bounds of Corollaries 4.3.2 and 4.3.5). Let $\alpha \in (0,1)$. Suppose that the assumptions of Corollaries 4.3.2 and 4.3.5 hold. Previous results show that each procedure may perform better than the other one in convenient regimes of the parameters, involving the unknown parameter s. In order to aggregate the two procedures into an automatic one, we take the maximum of the two test procedures. This aggregated test procedure rejects as soon as at least one of the procedures rejects, and accepts otherwise.

More precisely, let $\rho^{(1)}(\alpha/2)$ be defined by (4.41) with α replaced by $\alpha/2$ and set $t^{(1)} = (\rho^{(1)}(\alpha/2))^2/2$; and let $\rho^{(2)}(\alpha/2)$ be defined in (4.50) and $t^{(2)}$ be given by (4.51) with α replaced by $\alpha/2$. Then, Corollaries 4.3.2 and 4.3.5 imply that $R_{\rho^{(1)}}\left(\Psi_{\mathcal{T}_1}(t^{(1)})\right) \leq \alpha/2$ and $R_{\rho^{(2)}}\left(\Psi_{\mathcal{T}_2}(t^{(2)})\right) \leq \alpha/2$. We define the test:

$$\Psi^{\max} = \max(\Psi_{\mathcal{T}_1}(t^{(1)}), \Psi_{\mathcal{T}_2}(t^{(2)})).$$

It is straightforward to see that the type I error probability satisfies:

$$\sup_{(\beta^\star,\vartheta^\star)\in H_0} \mathbb{E}_{(\beta^\star,\vartheta^\star)}[\Psi^{\max}] \leq \alpha.$$

Moreover, we have for $\rho^{\min}(\alpha) = \rho^{(1)}(\alpha/2) \wedge \rho^{(2)}(\alpha/2)$ the following bound on the type II error probability:

$$\sup_{(\beta^\star,\vartheta^\star)\in H_1(\rho^{\min})} \mathbb{E}_{(\beta^\star,\vartheta^\star)}[1-\Psi^{\max}] \leq \alpha/2.$$

Therefore, we deduce an upper bound on $\rho^\star(\alpha)$ of order $\rho^{\min}(\alpha)$, that is:

$$\rho^{\min}(\alpha) = \min\left(\left(\frac{80\Xi_T}{\alpha}\right)^{1/4}, C\overline{\sigma}\sqrt{(s\vee s^0\vee 1)\Delta_T \log\left(\frac{2\,c\,|\Theta_T|}{\alpha\,\sigma_T}\right)}\right), \tag{4.52}$$

for a positive constant $c \geq 2$. We identify two regimes depending on whether the observed signal is sparse or not. Indeed, we notice that when α is fixed and:

$$s\vee s^0\vee 1 \ll \left(\frac{\Xi_T}{\alpha}\right)^{1/2} \cdot \left(\overline{\sigma}^2\Delta_T \log\left(\frac{2\,c\,|\Theta_T|}{\alpha\,\sigma_T}\right)\right)^{-1},$$

Corollary 4.3.5 yields a sharper upper bound on the separation rate than Corollary 4.3.2.

4.3.3 Minimax separation rates for signal detection

We illustrate our results on a simple model motivated by [Ingster et al., 2010] for sparse linear regression. We consider a discrete-time process y over a regular grid $t_1 < \cdots < t_T$ on $\Theta = \mathbb{R}/\mathbb{Z}$ with grid step $\Delta_T = 1/T$. We set λ_T and w_T as in Section 4.1.2.1. We recall that $\Xi_T = 2\overline{\sigma}^4\Delta_T^2 T$ where $\overline{\sigma} > 0$ is the noise level. In the following, we assume without any loss of generality that $\overline{\sigma} = 1$.

Let us consider the framework of signal detection when $s^0 = 0$. Under the assumptions of Corollary 4.3.5, the test problem (4.31) reduces to:

$$\begin{cases} H_0: & \beta^\star = 0, \\ H_1(\rho): & (\beta^\star,\vartheta^\star) \in (\mathbb{R}^*)^s \times \Theta_T^s(\delta^\star) \quad \text{such that} \quad \|\beta^\star\Phi_T(\vartheta^\star)\|_{L^2(\lambda_T)} \geq \rho. \end{cases} \tag{4.53}$$

Moreover, under the assumptions of Corollary 4.3.5 and with the same arguments used to establish (4.27), we can show that:

$$5/6 \leq C_{\min} := \min_\beta \frac{\|\beta\Phi_T(\vartheta^\star)\|_{L^2(\lambda_T)}}{\|\beta\|_{\ell_2}} \quad \text{and} \quad C_{\max} := \max_\beta \frac{\|\beta\Phi_T(\vartheta^\star)\|_{L^2(\lambda_T)}}{\|\beta\|_{\ell_2}} \leq 7/6.$$

Therefore, the separation in the alternative hypothesis $H_1(\rho)$ can be formulated as a lower bound on $\|\beta^\star\|_{\ell_2}$ since we have:

$$C_{\min}\|\beta^\star\|_{\ell_2} \leq \|\beta^\star \Phi_T(\vartheta^\star)\|_{L^2(\lambda_T)} \leq C_{\max}\|\beta^\star\|_{\ell_2}.$$

We set $\Theta_T = \Theta$ and thus $|\Theta_T| = 1$. We get from (4.52) the following upper bound on $\rho^\star(\alpha)$ for any $\alpha \in (0,1)$:

$$\rho(\alpha) = C \min\left(\frac{1}{(\alpha T)^{\frac{1}{4}}}, \sqrt{\frac{s}{T}\log\left(\frac{c}{\alpha\,\sigma_T}\right)}\right),$$

with C a finite positive constant. Let $(\alpha_T, T \geq 1)$ be a $(0,1)$-valued sequence which converges to zero when T grows to infinity. We deduce that:

$$\lim_{s,T\to+\infty} R^\star_{\rho(\alpha_T)} = 0.$$

By letting the sequence $(\alpha_T, T \geq 1)$ converge towards 0 as slow as we want, we deduce that for a sequence of separations $(\rho_{s,T},\, T \geq 1, s \geq 1)$ such that:

$$\lim_{s,T\to+\infty} \frac{\rho_{s,T}}{\frac{1}{T^{\frac{1}{4}}} \wedge \sqrt{\frac{s}{T}\log\left(\frac{c}{\sigma_T}\right)}} = +\infty,$$

we have:

$$\lim_{s,T\to+\infty} R^\star_{\rho_{s,T}} = 0.$$

Hence, we have obtained an asymptotic upper bound of the minimax separation associated to the detection of a mixture issued from a continuous dictionary.

We now compare this upper bound to the asymptotic lower bound obtained in the case where the dictionary contains a finite number of features instead of a continuum. Assume that the dictionary is fixed, known and contains p features parametrized by the parameters in the known and fixed set $\mathcal{Q}^0 = \{\theta_1^0, \cdots, \theta_p^0\} \subset \Theta_T$. We consider the high dimensional linear regression model:

$$y = \beta^\star \Phi_T(\vartheta^0) + w_T \quad \text{in } L^2(\lambda_T),$$

with $\vartheta^0 = (\theta_1^0, \cdots, \theta_p^0) \in \Theta_T^p$ and where $\beta^\star \in \mathbb{R}^p$ is a s-sparse vector. Notice that in this model the entries of $\beta^\star$ can take the value 0. The high dimension comes from the fact that p can be much larger than T. Under coherence assumptions on the finite dictionary and for a sequence of separations $(\rho_{s,T},\, T \geq 1, s \geq 1)$ such that:

$$\lim_{s,T\to+\infty} \frac{\rho_{s,T}}{\frac{1}{T^{\frac{1}{4}}} \wedge \sqrt{\frac{s}{T}\log(p)} \wedge \frac{p^{\frac{1}{4}}}{\sqrt{T}}} = 0, \tag{4.54}$$

the authors of [Ingster et al., 2010] showed for different hypotheses on the design matrix $\Phi_T(\vartheta^0)$ that:

$$\lim_{s,T\to+\infty} R^\star_{\rho_{s,T}} = 1.$$

It means that the hypotheses (4.53) cannot be distinguished asymptotically when the separation converges to zero faster than the rate given by (4.54). We remark that in the high dimensional framework (*i.e.* $T < p$), we get only the first two regimes in (4.54) since $1/T^{1/4} < p^{1/4}/\sqrt{T}$.

4.4 Goodness-of-fit of the dictionary

In spectroscopy, a prescribed material has known chemical components and a list of s_0 corresponding location parameters of the features is provided. From a sampled material we want to decide whether its chemical components are included in the prescribed list. The linear coefficients are non-negative in this case and they are not given, which makes the null hypothesis composite, that is, fixed location parameters and varying positive linear coefficients. We generalize this setup to real valued linear coefficients. Under the null hypothesis the location parameters are still fixed, but the linear coefficients vary with fixed sign.

More precisely, let $s^0 \in \mathbb{N}$ and let $\mathcal{Q}^0 = \{\theta_1^0, \cdots, \theta_{s_0}^0\} \subset \Theta_T$ be a set of known location parameters pairwise separated by a distance $\delta^0 \geq 0$ so that the model is identifiable, see Lemma 4.2.4. We set the vector $\vartheta^0 = (\theta_1^0, \cdots, \theta_{s^0}^0)$. Let $v^0 = (v_1^0, \ldots, v_{s^0}^0)$ be a vector in $\{-1, 1\}^{s^0}$ that contains the common signs of all linear coefficients under the null hypothesis. Consider two disjoint subsets of the set $\mathcal{Q}^0$ associated to linear coefficients with sign $\epsilon = \pm 1$: $\mathcal{Q}^{0,\epsilon} = \{\theta_k^0 \in \mathcal{Q}^0 : \epsilon v_k^0 > 0\}$. Let $s \in \mathbb{N}^*$. Assume that we observe a random element y issued from the model (4.2) with linear coefficients $\beta^\star \in (\mathbb{R}^*)^s$ and non-linear parameters $\vartheta^\star = (\theta_1^\star, \cdots, \theta_s^\star) \in \Theta_T^s$. We test if the unknown set $\mathcal{Q}^{\star,\epsilon} = \{\theta_k^\star \in \mathcal{Q}^\star : \epsilon\beta_k^\star > 0\}$ is a subset of $\mathcal{Q}^{0,\epsilon}$ for each $\epsilon = \pm 1$. If $s^0 = 0$, this amounts to testing that $\mathcal{Q}^\star$ is empty, which corresponds to the signal detection framework presented in Section 4.3 in the case $s^0 = 0$. Hence, we shall assume in this section that $s_0 \geq 1$. For example, if $\mathcal{Q}^{0,-}$ is empty, this amounts to testing that $\mathcal{Q}^\star$ is a subset of $\mathcal{Q}^0$ and $\beta^\star$ has positive entries.

4.4.1 A measure of discrepancy between dictionaries

We define the closed balls centered at $\theta \in \Theta_T$ with radius r by:

$$\mathcal{B}_T(\theta, r) = \{\theta' \in \Theta_T : \mathfrak{d}_T(\theta, \theta') \leq r\} \subseteq \Theta_T.$$

Let us define for $\epsilon = \pm 1$ the set of indices $\mathcal{I}^\epsilon = \{k \in \{1, \ldots, s^0\}, \epsilon v_k^0 > 0\}$. We introduce for $r > 0$, $k \in \mathcal{I}^\epsilon$ and $\epsilon \in \{-1, +1\}$ the set $S_k^\epsilon(r)$ gathering the indices of the elements of $\mathcal{Q}^{\star,\epsilon}$ that are close to the element θ_k^0 of $\mathcal{Q}^{0,\epsilon}$:

$$S_k^\epsilon(r) = \left\{\ell \in \{1, \cdots, s\} : \theta_\ell^\star \in \mathcal{B}_T(\theta_k^0, r) \text{ and } \epsilon\beta_\ell^\star > 0\right\}. \tag{4.55}$$

Notice that the sets $S_k^\epsilon(r)$ can be empty. We assume that $r < \min_{\ell \neq k} \mathfrak{d}_T(\theta_\ell^0, \theta_k^0)/2$ so that the sets $S_k^\epsilon(r)$ with $\epsilon = \pm 1$ and $k \in \mathcal{I}^\epsilon$ are pairwise disjoint. We also set:

$$S(r) = \bigcup_{\epsilon \in \{-1,+1\}} S^\epsilon(r) \quad \text{with} \quad S^\epsilon(r) = \bigcup_{k \in \mathcal{I}^\epsilon} S_k^\epsilon(r).$$

We now define a discrepancy measure between the model and any approximation by a linear combination of features having their parameters in $\mathcal{Q}^0$:

$$\mathcal{D}_{T,r}(\beta^\star, \vartheta^\star, v^0, \vartheta^0) = \sum_{\epsilon \in \{-1,+1\}} \sum_{k \in \mathcal{I}^\epsilon} \sum_{\ell \in S_k^\epsilon(r)} |\beta_\ell^\star| \, \mathfrak{d}_T(\theta_\ell^\star, \theta_k^0)^2 + \sum_{k \in S(r)^c} |\beta_k^\star| \quad \text{for } r > 0,$$

where $S(r)^c$ denotes the complementary set of $S(r)$ in $\{1, \ldots, s\}$. Notice that we have $\mathcal{D}_{T,r}(\beta^\star, \vartheta^\star, v^0, \vartheta^0) = 0$ if and only if $\mathcal{Q}^{\star,+} \subseteq \mathcal{Q}^{0,+}$ and $\mathcal{Q}^{\star,-} \subseteq \mathcal{Q}^{0,-}$.

4.4.2 The testing hypotheses

We shall test the following hypotheses:

$$\begin{cases} H_0 : & (\beta^\star, \vartheta^\star) \in (\mathbb{R}^*)^s \times \Theta_T^s(\delta^\star), \qquad \mathcal{Q}^{\star,+} \subseteq \mathcal{Q}^{0,+} \text{ and } \mathcal{Q}^{\star,-} \subseteq \mathcal{Q}^{0,-}, \\ H_1(\rho) : & (\beta^\star, \vartheta^\star) \in (\mathbb{R}^*)^s \times \Theta_T^s(\delta^\star) \quad \text{and} \quad \mathcal{D}_{T,r}(\beta^\star, \vartheta^\star, v^0, \vartheta^0) \geq \rho, \end{cases} \tag{4.56}$$

where ρ and $\delta^\star$ are separation parameters depending *a priori* on T, s and s^0 that need to be evaluated. Notice that the null hypothesis is also composite. We recall the definitions (4.29) and (4.30) of the minimax testing risk $R^\star_\rho$ and the minimax separation $\rho^\star$. In the following, we give upper bounds on the testing risk and on the minimax separation $\rho^\star(\alpha)$ for any $\alpha \in (0,1)$.

4.4.3 Main result

In this section, we build a test for (4.56). Under Assumptions 4.2.1 and 4.2.2, we define the element of $L^2(\lambda_T)$:

$$p_0 = \sum_{k=1}^{s^0} \alpha_k \phi_T(\theta_k^0) + \sum_{k=1}^{s^0} \xi_k \, \tilde{D}_{1,T}[\phi_T](\theta_k^0), \tag{4.57}$$

where $\alpha, \xi \in \mathbb{R}^{s^0}$ solve the system:

$$\Big\langle \phi_T(\theta_k^0), p_0 \Big\rangle_{L^2(\lambda_T)} = v_k^0 \quad \text{and} \quad \Big\langle \partial_\theta \phi_T(\theta_k^0), p_0 \Big\rangle_{L^2(\lambda_T)} = 0, \text{ for all } k \in \{1, \cdots, s^0\}. \tag{4.58}$$

Remark 4.4.1. The element p_0 of $L^2(\lambda_T)$ coincides with the vanishing derivative pre-certificate which appears in [Duval and Peyré, 2015, Section 4] and is the solution of (4.58) with minimal norm $\|p_0\|_{L^2(\lambda_T)}$.

Following [Butucea et al., 2022a], we give the existence and properties of the interpolating certificate function.

Lemma 4.4.2 (Interpolating certificate)**.** *Let $T \in \mathbb{N}$, let $s \in \mathbb{N}^*$, $r \in \big(0, 1/\sqrt{2\, g_\infty\, L_2}\big)$, $\eta \in (0,1)$ and $\mathcal{Q} = \{\theta_1, \cdots, \theta_s\} \subset \Theta_T$. Suppose that Assumption 4.2.4 holds.*

Then, there exist finite positive constants C_N, C_F, C_B with $C_F < 1$, depending on r and the function F, such that for any application $v : \mathcal{Q} \mapsto \{-1, 1\}$, there exist unique $\alpha, \xi \in \mathbb{R}^s$ such that $p \in L^2(\lambda_T)$ uniquely defined by:

$$\begin{cases} p = \sum\limits_{k=1}^{s} \alpha_k \phi_T(\theta_k) + \sum\limits_{k=1}^{s} \xi_k \, \tilde{D}_{1,T}[\phi_T](\theta_k), \\ \langle \phi_T(\theta), p \rangle_{L^2(\lambda_T)} = v(\theta) \quad \text{and} \quad \langle \partial_\theta \phi_T(\theta), p \rangle_{L^2(\lambda_T)} = 0, \quad \text{for all} \quad \theta \in \mathcal{Q}, \end{cases}$$

satisfies:

(i) For all $\theta \in \mathcal{Q}$ and $\theta' \in \mathcal{B}_T(\theta, r)$, we have $|\langle \phi_T(\theta'), p \rangle_{L^2(\lambda_T)}| \le 1 - C_N \, \mathfrak{d}_T(\theta, \theta')^2$.

(ii) For all θ in Θ_T, $\theta \notin \bigcup\limits_{\theta' \in \mathcal{Q}} \mathcal{B}_T(\theta', r)$ (far region), we have $|\langle \phi_T(\theta), p \rangle_{L^2(\lambda_T)}| \le 1 - C_F$.

(iii) We have $\|p\|_{L^2(\lambda_T)} \le \sqrt{s}\, C_B$.

Proof. Using similar arguments as those developed in the proof of Theorem 4.2.3, we get that all the hypotheses of [Butucea et al., 2022a, Proposition, 7.4] are satisfied. The existence and uniqueness of p is then guaranteed by [Butucea et al., 2022a, Lemma, 10.1]. The properties satisfied by p are direct consequences of [Butucea et al., 2022a, Proposition, 7.4]. □

Using the estimator $\hat{\beta}$ from (4.23) for a given value of the tuning parameter κ, we define the test statistic:

$$\mathcal{T}_3 = \left\| \hat{\beta} \right\|_{\ell_1} - \langle y, p_0 \rangle_{L^2(\lambda_T)}. \tag{4.59}$$

and the corresponding test $\Psi_{\mathcal{T}_3}(t) = \mathbf{1}_{\{|\mathcal{T}_3| > t\}}$.

Theorem 4.4.3. *Let $T \in \mathbb{N}$, $s^0 \in \mathbb{N}^*$ and choose $K \in \mathbb{N}$ such that $s_0 \le K$. Let also $\eta \in (0,1)$ and $r \in \big(0, 1/\sqrt{2\, g_\infty\, L_2}\big)$. Let $\delta^\star \ge \sigma_T \, \Sigma(\eta, r, s)$ and $\delta^0 \ge \sigma_T \, \Sigma(\eta, r, s^0)$. Assume we observe the random element y of $L^2(\lambda_T)$ under the regression model (4.2) with unknown parameters*

$s \in \mathbb{N}^*$ such that $s \leq K$, $\beta^\star \in (\mathbb{R}^*)^s$ and $\vartheta^\star = (\theta_1^\star, \cdots, \theta_s^\star) \in \Theta_T^s(\delta^\star)$. Let $v^0 \in \{-1,1\}^{s^0}$ be a sign vector and let $\vartheta^0 = (\theta_1^0, \cdots, \theta_{s^0}^0) \in \Theta_T^{s^0}(\delta^0)$. Suppose that Assumption 4.2.4 holds for the sets $\mathcal{Q}^\star = \{\theta_1^\star, \cdots, \theta_s^\star\} \subset \Theta_T$ of cardinal s and $\mathcal{Q}^0 = \{\theta_1^0, \cdots, \theta_{s^0}^0\} \subset \Theta_T$ of cardinal s^0. Suppose also that the noise process w_T satisfies Assumption 4.1.1 for a noise level $\overline{\sigma} > 0$ and a decay rate for the noise variance $\Delta_T > 0$.

Then, the test statistic $\mathcal{T}_3$ is uniquely defined and there exist finite positive constants, a and $\mathcal{C}_i$ with $i = 1, \cdots, 5$, (depending on r and on the function F) such that for any $\tau > 1$ and any tuning parameter κ:

$$\kappa \geq \mathcal{C}_1 \overline{\sigma} \sqrt{\Delta_T \log(\tau)}, \tag{4.60}$$

the test $\Psi_{\mathcal{T}_3}$ satisfies:

$$R_\rho\left(\Psi_{\mathcal{T}_3}(t)\right) \leq 2\,\mathcal{C}_2 \left(\frac{|\Theta_T|}{\sigma_T\, \tau\, \sqrt{\log(\tau)}} \vee \frac{1}{\tau} \right) + \frac{2}{\tau^{a\, s_0}}, \tag{4.61}$$

for any threshold $t > 0$ and any separation $\rho > 0$ satisfying:

$$t \geq 2\,\mathcal{C}_3\, s^0\, \kappa \quad \text{and} \quad \rho \geq \mathcal{C}_4\, s\, \kappa + \mathcal{C}_5\, t.$$

Proof. Recall the test problem given by (4.56). Assumption 4.2.4 holds for the set $\mathcal{Q}^0$. Thanks to Lemma 4.4.2, the element p_0 of $L^2(\lambda_T)$ is uniquely defined by v^0, (4.57) and (4.58). Hence, the test statistic $\mathcal{T}_3$ from (4.59) is well-defined.

We first bound the type I error probability. Let us fix $(\beta^\star, \vartheta^\star) \in (\mathbb{R}^*)^s \times \Theta_T^s(\delta^\star)$ such that H_0 holds. Using that $y = \beta^\star \Phi_T(\vartheta^\star) + w_T$ and the triangle inequality, we obtain:

$$\begin{aligned} |\mathcal{T}_3| &= \left| \left\|\hat{\beta}\right\|_{\ell_1} - \|\beta^\star\|_{\ell_1} + \|\beta^\star\|_{\ell_1} - \langle \beta^\star \Phi_T(\vartheta^\star), p_0 \rangle_{L^2(\lambda_T)} - \langle w_T, p_0 \rangle_{L^2(\lambda_T)} \right| \\ &\leq \left| \left\|\hat{\beta}\right\|_{\ell_1} - \|\beta^\star\|_{\ell_1} \right| + |B| + \left| \langle w_T, p_0 \rangle_{L^2(\lambda_T)} \right|, \end{aligned} \tag{4.62}$$

where:

$$B = \|\beta^\star\|_{\ell_1} - \langle \beta^\star \Phi_T(\vartheta^\star), p_0 \rangle_{L^2(\lambda_T)}\,. \tag{4.63}$$

Since $\mathcal{Q}^{\star,+} \subseteq \mathcal{Q}^{0,+}$, $\mathcal{Q}^{\star,-} \subseteq \mathcal{Q}^{0,-}$, we have for all $k \in \{1, \cdots, s\}$:

$$|\beta_k^\star| - \langle \beta_k^\star \phi_T(\theta_k^\star), p_0 \rangle_{L^2(\lambda_T)} = 0,$$

we deduce that $B = 0$ under H_0. Hence, we have that:

$$\mathbb{E}_{(\beta^\star, \vartheta^\star)}[\Psi_{\mathcal{T}_3}(t)] \leq \mathbb{P}\left(\left| \left\|\hat{\beta}\right\|_{\ell_1} - \|\beta^\star\|_{\ell_1} \right| > t/2 \right) + \mathbb{P}\left(\left| \langle w_T, p_0 \rangle_{L^2(\lambda_T)} \right| > t/2 \right). \tag{4.64}$$

Recall that under H_0, we have $s \leq s^0$. Therefore, since $\mathcal{C}_3\, \kappa\, s^0 \leq t/2$, we have $\mathcal{C}_3\, \kappa\, s \leq t/2$. We get from Theorem 4.2.3 that:

$$\mathbb{P}\left(\left| \left\|\hat{\beta}\right\|_{\ell_1} - \|\beta^\star\|_{\ell_1} \right| > t/2 \right) \leq \mathcal{C}_2 \left(\frac{|\Theta_T|}{\sigma_T \tau \sqrt{\log(\tau)}} \vee \frac{1}{\tau} \right). \tag{4.65}$$

Then, thanks to Assumptions 4.1.1 and Lemma 4.4.2, the quantity $\langle w_T, p_0 \rangle_{L^2(\lambda_T)}$ is a centered Gaussian random variable of variance bounded by $\overline{\sigma}^2 C_B^2 \Delta_T s_0$ where C_B is the finite positive constant from Lemma 4.4.2. Hence we have, provided that $t \geq 2\mathcal{C}_3\, \kappa\, s^0$ with $\kappa \geq \mathcal{C}_1 \overline{\sigma} \sqrt{\Delta_T \log(\tau)}$, that is, $t^2 \geq (2\mathcal{C}_1 \mathcal{C}_3 \overline{\sigma} s_0)^2 \Delta_T \log(\tau)$:

$$\mathbb{P}\left(\langle w_T, p_0 \rangle_{L^2(\lambda_T)} > t/2 \right) \leq \int_{t/2}^{+\infty} \frac{\mathrm{e}^{-x^2/(2\overline{\sigma}^2 \Delta_T C_B^2 s_0)}}{\sqrt{2\pi \overline{\sigma}^2 \Delta_T C_B^2 s_0}}\, \mathrm{d}x \leq \frac{1}{2}\, \mathrm{e}^{-\frac{t^2}{8(\overline{\sigma}^2 \Delta_T C_B^2 s_0)}} \leq \frac{1}{2\tau^{a s_0}},$$

with $a = (\mathcal{C}_1\mathcal{C}_3/C_B)^2/2$ and where we used the tail bound (4.39). It gives by symmetry that:

$$\mathbb{P}\left(|\,\langle w_T, p_0\rangle_{L^2(\lambda_T)}\,| > t/2\right) \leq \frac{1}{\tau^{a\,s_0}}. \tag{4.66}$$

Plugging (4.65) and (4.66) in (4.64), we get:

$$\sup_{(\beta^\star,\,\vartheta^\star)\in H_0} \mathbb{E}_{(\beta^\star,\,\vartheta^\star)}\left[\Psi_{\mathcal{T}_3}(t)\right] \leq \mathcal{C}_2\left(\frac{|\Theta_T|}{\sigma_T\,\tau\,\sqrt{\log(\tau)}} \vee \frac{1}{\tau}\right) + \frac{1}{\tau^{a\,s_0}}. \tag{4.67}$$

We now bound the type II error probability. Assume that H_1 holds, in this case we have $\mathcal{D}_{T,r}(\beta^\star, \vartheta^\star, v^0, \vartheta^0) \geq \rho$. We have, using the first equality of (4.62) and the triangle inequality, that:

$$|\mathcal{T}_3| \geq |B| - \left|\langle w_T, p_0\rangle_{L^2(\lambda_T)}\right| - \left|\left\|\hat{\beta}\right\|_{\ell_1} - \|\beta^\star\|_{\ell_1}\right|,$$

with B defined in (4.63). Using the definitions (4.55) of $S(r)$ and $S_k^\epsilon(r)$ with $\epsilon \in \{-1,+1\}$ and $k \in \mathcal{I}^\epsilon$, we get:

$$B = \sum_{\substack{\epsilon\in\{-1,+1\}\\ k\in\mathcal{I}^\epsilon,\,\ell\in S_k^\epsilon(r)}} |\beta_\ell^\star|\left(1 - \mathrm{sign}(\beta_\ell^\star)\,\langle \phi_T(\theta_\ell^\star), p_0\rangle_{L^2(\lambda_T)}\right) + \sum_{k\in S(r)^c} |\beta_k^\star|\left(1 - \mathrm{sign}(\beta_k^\star)\,\langle \phi_T(\theta_k^\star), p_0\rangle_{L^2(\lambda_T)}\right).$$

Thanks to Lemma 4.4.2 (i)-(ii) of , we obtain:

$$\begin{aligned} B &\geq \sum_{\substack{\epsilon\in\{-1,+1\}\\ k\in\mathcal{I}^\epsilon,\,\ell\in S_k^\epsilon(r)}} C_N|\beta_\ell^\star|\mathfrak{d}_T(\theta_\ell^\star, \theta_k^0)^2 + \sum_{k\in S(r)^c} C_F|\beta_k^\star| \\ &\geq (C_N \wedge C_F)\mathcal{D}_{T,r}(\beta^\star, \vartheta^\star, v^0, \vartheta^0) \geq (C_N \wedge C_F)\rho, \end{aligned}$$

where the constants C_N and C_F are defined in Lemma 4.4.2 and depend on r and on the function F. Therefore, we have with $a_t = (C_N \wedge C_F)\rho - t$:

$$\begin{aligned} \mathbb{E}_{(\beta^\star,\vartheta^\star)}\left[1 - \Psi_{\mathcal{T}_3}(t)\right] &\leq \mathbb{P}\left(\left|\langle w_T, p_0\rangle_{L^2(\lambda_T)}\right| + \left|\|\beta^\star\|_{\ell_1} - \left\|\hat{\beta}\right\|_{\ell_1}\right| \geq a_t\right) \\ &\leq \mathbb{P}\left(\left|\langle w_T, p_0\rangle_{L^2(\lambda_T)}\right| \geq a_t/2\right) + \mathbb{P}\left(\left|\|\beta^\star\|_{\ell_1} - \left\|\hat{\beta}\right\|_{\ell_1}\right| \geq a_t/2\right). \end{aligned}$$

Provided that $\rho \geq \mathcal{C}_4\,s\,\kappa + \ \mathcal{C}_5\,t$ with $\mathcal{C}_4 = 2\,\mathcal{C}_3/(C_N \wedge C_F)$ and $\mathcal{C}_5 = 2/(C_N \wedge C_F)$ we have $a_t/2 \geq (\mathcal{C}_3\kappa s) \vee (t/2)$. By using (4.65) and (4.66), we obtain:

$$\sup_{(\beta^\star,\,\vartheta^\star)\in H_1(\rho)} \mathbb{E}_{(\beta^\star,\,\vartheta^\star)}\left[1 - \Psi_{\mathcal{T}_3}(t)\right] \leq \mathcal{C}_2\left(\frac{|\Theta_T|}{\sigma_T\,\tau\,\sqrt{\log(\tau)}} \vee \frac{1}{\tau}\right) + \frac{1}{\tau^{a s_0}}. \tag{4.68}$$

Finally, by adding both sides of (4.67) and (4.68), we get (4.61). This concludes the proof. □

4.4.4 Separation rates

We give in this section an upper bound on the minimax separation $\rho^\star$ to test the goodness-of-fit of the dictionary, that is to distinguish the assumptions H_0 and $H_1(\rho)$ presented in Section 4.4.

Theorem 4.4.4. *Under the framework and the assumptions of Theorem 4.4.3, there exist finite positive constants c and C (depending on r and the function F) such that provided that $|\Theta_T|/\sigma_T \geq 1$, we have for any $\alpha \in (0,1)$:*

$$\rho^\star(\alpha) \leq C\,\overline{\sigma}\,(s \vee s^0)\sqrt{\Delta_T \log\left(\frac{c\,|\Theta_T|}{\alpha\,\sigma_T}\right)}. \tag{4.69}$$

Proof. The result is a direct consequence of Theorem 4.4.3. We fix the tuning parameter $\kappa = \mathcal{C}_1\overline{\sigma}\sqrt{\Delta_T \log(\tau)}$ by taking the equality in (4.60). Then, for $\rho \geq \mathcal{C}_4\, s\, \kappa + \mathcal{C}_5\, t$ and $t = 2\,\mathcal{C}_3\, s^0\kappa$ we have by Theorem 4.4.3 for $\tau > 1$ and since $s_0 \geq 1$:

$$R^\star_\rho \leq R_\rho\left(\Psi_{\mathcal{T}_3}(t)\right) \leq 2\mathcal{C}_2\left(\frac{|\Theta_T|}{\sigma_T\,\tau\,\sqrt{\log(\tau)}} \vee \frac{1}{\tau}\right) + \frac{2}{\tau^a},$$

where the finite positive constants a, $\mathcal{C}_i$ with $i \in \{1, \cdots, 5\}$, from Theorem 4.4.3 depend on r and the function F.

Hence, by taking $\tau = c'/(\sigma_T\alpha/(2|\Theta_T|))^{c''}$ with $c'' = 1 \vee (1/a)$ and $c' = (2\mathcal{C}_2) \vee \mathrm{e} \vee 2^{1/a}$, we get for $\rho \geq 2\mathcal{C}_1((2\,\mathcal{C}_3\,\mathcal{C}_5) \vee \mathcal{C}_4)\overline{\sigma}(s \vee s^0)\sqrt{\Delta_T \log(c'/(\sigma_T\alpha/(2|\Theta_T|))^{c''})}$ and $\alpha \in (0,1)$ that $R^\star_\rho \leq \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$. We readily deduce (4.69) with $c = 2c'^{(1/c'')}$. □

4.5 Gaussian scaled spikes deconvolution

In this section, we consider the discrete time process observed on a regular grid given in Section 4.1.2.1. We recall that Assumption 4.1.1 holds with:

$$\lambda_T = \Delta_T \sum_{j=1}^{T} \delta_{t_j} \quad \text{with} \quad t_j = -b_T + j\Delta_T \quad \text{and} \quad \Delta_T = \frac{2b_T}{T},$$

and w_T given by (4.4), where $T \in \mathbb{N}^*$. We consider the scaled Gaussian features associated to the function:

$$h(t,\sigma) \mapsto \frac{\exp(-t^2/2\sigma^2)}{\pi^{1/4}\sigma^{1/2}} \quad \text{defined on} \quad \Theta \times \mathfrak{S} = \mathbb{R} \times \mathbb{R}_+^*.$$

We shall see below that the natural choice for the function F appearing in (4.9) is given by:

$$F = h^0 * h^0 = \pi^{1/4} h^0(\cdot/\sqrt{2}) \quad \text{with} \quad h^0(\cdot) = h(\cdot, 1).$$

In the following, we check that Assumption 4.2.4 holds. Then, using Theorem 4.2.3 on a particular example, we provide a prediction bound for the estimator of $(\beta^\star, \vartheta^\star)$ solution of the optimization problem (4.23).

4.5.1 Choice of the approximating kernel

We denote the unscaled feature φ^0 on $\theta \in \Theta$ by:

$$\varphi^0(\theta) = h(\theta - \cdot, 1) = h^0(\theta - \cdot).$$

We define the mapping $f_T : \Theta \to \Theta$ by $f_T(\theta) = \theta/\sigma_T$ for any $\theta \in \Theta$ and the (pushforward) measure $\lambda^0_T = \lambda_T \circ f_T^{-1}$ so that for any $g \in L^1(\lambda^0_T)$:

$$\int g(\theta/\sigma_T)\lambda_T(\mathrm{d}\theta) = \int g(\theta)\,\lambda^0_T(\mathrm{d}\theta).$$

The Hilbert space $L^2(\lambda^0_T)$ is endowed with its natural scalar product $\langle\cdot,\cdot\rangle_{L^2(\lambda^0_T)}$ and norm $\|\cdot\|_{L^2(\lambda^0_T)}$. We define on Θ^2 the kernel:

$$\mathcal{K}^0_T(\theta,\theta') = \langle \phi^0_T(\theta), \phi^0_T(\theta')\rangle_{L^2(\lambda^0_T)} \quad \text{with} \quad \phi^0_T(\theta) = \varphi^0(\theta)/\left\|\varphi^0(\theta)\right\|_{L^2(\lambda^0_T)}.$$

The kernel $\mathcal{K}_T$ can be seen as a scaled kernel derived from $\mathcal{K}^0_T$ as for any $\theta, \theta' \in \Theta$:

$$\mathcal{K}_T(\theta,\theta') = \mathcal{K}^0_T(\theta/\sigma_T, \theta'/\sigma_T).$$

When the measure λ_T^0 converges in some sense, as T goes to infinity, towards the Lebesgue measure Leb on $\mathbb{R}$, it is natural to consider the approximation $\mathcal{K}_\infty^0$ of $\mathcal{K}_T^0$ on Θ^2 by:

$$\mathcal{K}_\infty^0(\theta,\theta') = \left\langle \phi_\infty^0(\theta), \phi_\infty^0(\theta') \right\rangle_{L^2(\mathrm{Leb})} \quad \text{with} \quad \phi_\infty^0(\theta) = \varphi^0(\theta) / \left\| \varphi^0(\theta) \right\|_{L^2(\mathrm{Leb})}.$$

Thanks to the definition of F, we also have on Θ^2 that:

$$F(\theta - \theta') = \mathcal{K}_\infty^0(\theta,\theta').$$

The approximating kernel $\mathcal{K}_T^{\mathrm{prox}}$ is then given by (4.9) on Θ^2.

4.5.2 Checking Assumption 4.2.4

4.5.2.1 Regularity of the dictionary

We refer to [Butucea et al., 2022a, Section 8] to check that Assumption 4.2.4 *(i)* holds for the feature φ_T defined by (4.1) and any scale parameter $\sigma_T \in \mathfrak{S} = \mathbb{R}_+^*$.

4.5.2.2 Boundedness and local concavity on the diagonal

Elementary calculations show that $g_\infty = -F''(0) = 1/2$. By definition of F, we directly deduce that Assumption 4.2.3 holds. We also get that for $r \in (0, \sqrt{2})$:

$$\varepsilon(r) = 1 - \mathrm{e}^{-r^2/4} > 0 \quad \text{and} \quad \nu(r) = \left(1 - \frac{r^2}{2}\right) \mathrm{e}^{-r^2/4}.$$

We fix $r \in (0, 1/2)$. We readily check that Assumption 4.2.4 *(ii)* is verified.

4.5.2.3 Proximity to the approximating kernel

In order for the kernel $\mathcal{K}_T^{\mathrm{prox}}$ to be a good approximation of $\mathcal{K}_T$ in the sense of Assumption 4.2.4 *(iii)*, we shall consider the set Θ_T over which the optimization is performed:

$$\Theta_T = [(1-\xi)a_T, (1-\xi)b_T] \subset [a_T, b_T] \quad \text{with a given shrinkage parameter } \xi \in (0,1).$$

Intuitively, one does not expect the estimation of the location parameter to perform well near the lower and upper bounds of the observation grid (given by the support of λ_T). Following [Butucea et al., 2022a, Section 8], we set:

$$\gamma_T = 2\Delta_T \, \sigma_T^{-1} + \sqrt{\pi} \, \mathrm{e}^{-\xi^2 b_T^2 / 2\sigma_T^2}. \tag{4.70}$$

Recall $\mathcal{V}_T$ and C_T defined by (4.17) and (4.19). Using Lemma [Butucea et al., 2022a, Lemma 8.1], there exist finite positive universal constants c_0, c_1 and c_2, such that $\gamma_T < c_0$ implies:

$$\mathcal{V}_T \le c_1 \gamma_T \quad \text{and} \quad |1 - C_T| \le c_2 \gamma_T.$$

Assume that $(b_T, T \ge 2)$ and $(\sigma_T, T \ge 2)$ are sequences of positive numbers, such that:

$$\lim_{T\to\infty} b_T = +\infty, \quad \lim_{T\to\infty} \sigma_T = 0 \quad \text{and} \quad \lim_{T\to\infty} \Delta_T \, \sigma_T^{-1} = 0. \tag{4.71}$$

Therefore, we have $\lim_{T\to+\infty} \mathcal{V}_T = 0$ and $\lim_{T\to+\infty} C_T = 1$.

Let $\eta \in (0,1)$ be fixed. We deduce that under (4.71), Assumption 4.2.4 *(iii)* is satified provided that T is larger than some constant depending on η, r, the sparsity s and the sequences $(b_T, T \ge 2)$ and $(\sigma_T, T \ge 2)$.

4.5.2.4 Separation of the non-linear parameters

We remark that $\lim_{r''\to\infty} \sup_{|r'|\geq r''} |F^{(i)}(r')| = 0$ for all $i \in \{0,\dots,3\}$. Thus, we deduce from the definition (4.20) of δ that $\delta(u,s)$ is finite for all $s \in \mathbb{N}^*$ and $u > 0$. Let us stress that $\sup_{s\in\mathbb{N}^*} \delta(u,s) \leq M/u$ for some universal finite constant M, see [Butucea et al., 2022a, Remark 8.2]. Therefore, the quantity $\Sigma(\eta,r,s)$ is bounded by a constant depending only on η and r.

So Assumption 4.2.4 *(iv)* is verified as soon as $|\theta-\theta'| > \sigma_T\,\Sigma(\eta,r,s)$ for all for all $\theta \neq \theta' \in \mathcal{Q}^\star$. (Notice this happens for the scaling parameter σ_T small enough depending on $\mathcal{Q}^\star$.)

4.5.3 Prediction error bound in a particular case

Recall the shrinkage parameter $\xi \in (0,1)$ in (4.70). Let us assume that:

$$b_T = \log(T) \quad \text{and} \quad \sigma_T = 1/\sqrt{\xi\log(T)}.$$

In particular, condition (4.71) holds. In this case, there exists a finite positive constant c depending on r, η and ξ such that for $T \geq c\log(T)^{3/2}\,s$, Assumption 4.2.4 holds (notice that the separation condition (4.22) of the location parameters in $\mathcal{Q}^\star$ is also verified for T large enough, depending on $\mathcal{Q}^\star$, as $\lim_{T\to+\infty}\sigma_T = 0$). By Theorem 4.2.3 with $\tau = T$ and κ given by the equality in (4.24), we get that:

$$\frac{1}{\sqrt{T}}\left\|\hat\beta\Phi_T(\hat\vartheta) - \beta^\star\Phi_T(\vartheta^\star)\right\|_{\ell_2} \leq \mathcal{C}_0\,\mathcal{C}_1\,\overline{\sigma}\sqrt{\frac{s\log(T)}{T}},$$

with probability larger than $1 - \mathcal{C}_2\left(\frac{2\sqrt{\xi}\log(T)}{T} \vee \frac{1}{T}\right)$, where the constants $\mathcal{C}_0$, $\mathcal{C}_1$ and $\mathcal{C}_2$ do not depend on T.

4.6 Low-pass filter

In this section, we consider the continuous-time process described in Section 4.1.2.2 on the torus $\Theta = \mathbb{R}/\mathbb{Z}$ with $\lambda_T = \mathrm{Leb}$ the Lebesgue measure on Θ and the noise:

$$w_T = \sum_{k\in\mathbb{N}} \sqrt{\xi_k}\,G_k\,\psi_k,$$

where $(G_k, k\in\mathbb{N})$ are independent centered Gaussian random variables with variance $\overline{\sigma}^2$, $(\psi_k, k\in\mathbb{N})$ is an o.n.b. of $L^2(\mathrm{Leb})$ on Θ and $\xi = (\xi_k, k\in\mathbb{N})$ is a square summable sequence of non-negative real numbers depending on $T \in 2\mathbb{N}^*+1$. Recall that the noise satisfies Assumption 4.1.1 for a positive noise level $\overline{\sigma}$ and a decay on the noise variance $\Delta_T = \sup_{k\in\mathbb{N}}\xi_k$.

We consider the normalized Dirichlet kernel, see (4.7), on Θ:

$$h(t,\sigma) = \frac{\sin(T\pi t)}{\sqrt{T}\,\sin(\pi t)} \quad \text{defined for} \quad t\in\Theta=\mathbb{R}/\mathbb{Z} \quad \text{and} \quad \sigma = \frac{1}{T}, \quad T\in 2\mathbb{N}^*+1.$$

The parameter T is related to the so-called cut-off frequency $f_c \in \mathbb{N}^*$ by $T = 2f_c+1$. We shall see below that the natural choice for the function F appearing in (4.9) is given by:

$$F(t) = \frac{\sin(\pi t)}{\pi t} \quad \text{for} \quad t\in\mathbb{R}.$$

We get from the definition (4.12) that $g_\infty = -F''(0) = \pi^2/3$.

In the following, we check that Assumption 4.2.4 hold. Then, using Theorem 4.2.3, we provide a prediction bound for the estimator of $(\beta^\star,\vartheta^\star)$ solution of the optimization problem (4.23).

4.6.1 The approximating kernel

We define the features φ_T using (4.1) with $\sigma_T = 1/T$. Elementary calculations give that for $\theta, \theta' \in \Theta$:

$$\mathcal{K}_T(\theta, \theta') = \frac{\sin(T\pi(\theta - \theta'))}{T \sin(\pi(\theta - \theta'))}.$$

Recall that by convention $|\theta - \theta'|$ is the Euclidean distance between θ and θ' in Θ, and in particular it belongs to $[0, 1/2]$. We define the approximating kernel $\mathcal{K}_T^{\mathrm{prox}}$ on Θ by:

$$\mathcal{K}_T^{\mathrm{prox}}(\theta, \theta') = F(T|\theta - \theta'|) \quad \text{with} \quad |\theta - \theta'| \in [0, 1/2].$$

Since F is even, we get also that $F(T|\theta - \theta'|) = F(T(\theta - \theta'))$ where, for $\theta, \theta' \in \Theta$, their representers in $\mathbb{R}$ a chosen so that $\theta - \theta'$ belongs to $[-1/2, 1/2]$.

4.6.2 Checking Assumption 4.2.4

4.6.2.1 Regularity of the dictionary

It is elementary to check that g_T is a constant function on Θ equal to $g_\infty\,(T^2 - 1)$ and that Assumption 4.2.4 *(i)* on the regularity of the dictionary holds.

4.6.2.2 Boundedness and local concavity on the diagonal

There exists $R > 0$ such that for any $r \in (0, R)$:

$$\varepsilon(r) = 1 - \frac{\sin(\pi r)}{\pi r} > 0 \quad \text{and} \quad \nu(r) = -\left(\frac{6}{\pi^3 r^3} - \frac{3}{\pi\, r}\right)\sin(\pi r) + \frac{6\cos(\pi r)}{\pi^2 r^2} > 0.$$

We fix $r \in (0, (1/\sqrt{2 g_\infty L_2}) \wedge (R/2))$. This and the fact that F is $\mathcal{C}^\infty$ with bounded derivatives implies that Assumption 4.2.4 *(ii)* on the boundedness and the local concavity of the approximating kernel holds.

4.6.2.3 Proximity to the approximating kernel

We set $\Theta_T = \Theta$. The proof of the next lemma on the uniform approximation of $\mathcal{K}_T$ by $\mathcal{K}_T^{\mathrm{prox}}$ on the torus is postponed to Section 4.6.3.1.

Lemma 4.6.1. *There exists a universal positive finite constant c_3 such that for any $T \in 2\mathbb{N}^* + 1$:*

$$\mathcal{V}_T \le \frac{c_3}{T} \quad \textit{and} \quad |1 - C_T| \le \frac{1}{2(T^2 - 1)}. \tag{4.72}$$

Let $\eta \in (0, 1)$ be fixed. We deduce from (4.72) that Assumption 4.2.4 *(iii)* is satified provided that T is larger than some constant depending on η, r, the sparsity s.

4.6.2.4 Separation of the non-linear parameters

Notice that $\lim_{r'' \to \infty} \sup_{|r'| \ge r''} |F^{(i)}(r')| = 0$ for all $i \in \{0, \cdots, 3\}$. Thus, we deduce from the definition (4.20) of δ that $\delta(u, s)$ is finite for all $s \in \mathbb{N}^*$ and $u > 0$.

So Assumption 4.2.4 *(iv)* is verified as soon as $|\theta - \theta'| > \sigma_T\, \Sigma(\eta, r, s)$ for all $\theta \neq \theta' \in \mathcal{Q}^\star$. (Notice this happens for T large enough depending on $\mathcal{Q}^\star$ as $\sigma_T = 1/T$.)

4.6.3 Prediction error bound

There exists a constant c depending on η and r such that for any $T \in 2\mathbb{N}^* + 1$ such that $T \geq c\, s$, and provided that (4.22) is satisfied, Assumption 4.2.4 holds. Using Theorem 4.2.3 with κ given by an equality in (4.24) with $\tau > 1$, we obtain the prediction bound:

$$\left\| \hat{\beta} \Phi_T(\hat{\vartheta}) - \beta^\star \Phi_T(\vartheta^\star) \right\|_{L^2(\mathrm{Leb})} \leq \mathcal{C}_0\, \mathcal{C}_1\, \overline{\sigma}\, \sqrt{s\, \Delta_T\, \log(\tau)},$$

with probability larger than $1 - \mathcal{C}_2 \left(\frac{T}{\tau \sqrt{\log(\tau)}} \vee \frac{1}{\tau} \right)$, where the constants $\mathcal{C}_0$, $\mathcal{C}_1$ and $\mathcal{C}_2$ do not depend on T.

Remark 4.6.2. Exact support recovery results were obtained in [Duval and Peyré, 2015]. The authors considered a small noise regime , that is $\|w_T\|_{L^2(\mathrm{Leb})}/\kappa$ less than a constant). They assumed that the location parameters satisfy for any $k, \ell \in \{1, \cdots, s\}$ such that $k \neq \ell$, the separation condition $|\theta_k^\star - \theta_\ell^\star| \geq C/f_c$ for $T = 2f_c + 1$, for some positive constant C and with $f_c \geq s$ (s being the number of active features in the mixture). They showed that there exist finite constants C' and C'' such that for all $k \in \{1, \cdots, s\}$:

$$|\tilde{\theta}_k - \theta_k^\star| \leq C' \|w_T\|_{L^2(\mathrm{Leb})} \quad \text{and} \quad |\tilde{\beta}_k - \beta_k^\star| \leq C'' \|w_T\|_{L^2(\mathrm{Leb})},$$

for some estimators $(\tilde{\beta}, \tilde{\vartheta} = (\tilde{\theta}_1, \cdots, \tilde{\theta}_s))$ obtained by solving the BLasso problem.

However the small noise regime assumption is restrictive as it does not encompass the example of Section 4.1.2.2 where for all $k \in \mathbb{N}$, $\xi_k = T^{-1} \mathbf{1}_{\{1 \leq k \leq T\}}$ and thus $\Delta_T = 1/T$ and $\mathbb{E}[\|w_T\|_{L^2(\mathrm{Leb})}]$ is of order 1. Recall that in (4.26) we obtain that our estimators satisfy:

$$\left| \|\hat{\beta}\|_{\ell_1} - \|\beta^\star\|_{\ell_1} \right| \leq C \frac{s\, \sqrt{\log(T)}}{\sqrt{T}}$$

for some constant $C > 0$ with high probability. Thus our prediction and estimation rates are smaller by a factor $\sqrt{\log(T)}/\sqrt{T}$ due to the probabilistic bounds on linear functionals of the noise process that we used in the proof, and this holds under analogous separation condition on any $\theta_k^\star$ and $\theta_\ell^\star$, for $k \neq \ell$ in $\{1, ..., s\}$.

4.6.3.1 Proof of Lemma 4.6.1

It is easy to check that the functions g_T and $g_{\mathcal{K}_T^{\mathrm{prox}}}$ are constant functions with:

$$g_T = g_\infty\, (T^2 - 1) \quad \text{and} \quad g_{\mathcal{K}_T^{\mathrm{prox}}} = g_\infty\, T^2.$$

Thus, we easily deduce the second inequality of (4.72) from the definition (4.17) of C_T.

We now consider the bound on $\mathcal{V}_T$. For $i, j \in \{0, \cdots, 3\}$ and $\ell = i + j$, we have with $\alpha_T = 1 - 1/T^2$:

$$\sup_{\Theta^2} |\mathcal{K}_T^{[i,j]} - \mathcal{K}_T^{\mathrm{prox}[i,j]}| = g_\infty^{-\ell/2} (T^2 \alpha_T)^{-\ell/2} \sup_{t \in [-\frac{1}{2}, \frac{1}{2}]} \left| \partial_t^\ell \left[D_T(t) + \left(1 - \alpha_T^{\ell/2}\right) \frac{\sin(T\pi t)}{T\pi t} \right] \right|, \quad (4.73)$$

where, for $t \in [-1/2, 1/2]$ and the convention $J(0) = 0$:

$$D_T(t) = \frac{\sin(T\pi t)}{T} J(t) \quad \text{and} \quad J(t) = \frac{1}{\sin(\pi t)} - \frac{1}{\pi t}.$$

It is easy to check that the function J can be expanded as a power series at 0 with positive convergence radius, and thus is of class $\mathcal{C}^\infty$ on $[-1/2, 1/2]$. Thus the following constant is finite:

$$M = \sup_{0 \leq \ell \leq 6} \sup_{[-1/2, 1/2]} |J^{(\ell)}| < +\infty.$$

Using the Leibniz rule, we have that for $\ell \in \{1, \cdots, 6\}$ and $t \in [-1/2, 1/2]$:

$$|\partial_t^\ell D_T(t)| = \frac{1}{T}\left|\sum_{j=0}^{\ell}\binom{\ell}{j}(T\pi)^j \sin^{(j)}(T\pi t)\, J^{(\ell-j)}(t)\right| \leq M\frac{(T\pi+1)^\ell}{T}.$$

We deduce from (4.73) that for $i, j \in \{0, \cdots, 3\}$ and $\ell = i + j$:

$$\sup_{\Theta^2}|\mathcal{K}_T^{[i,j]} - \mathcal{K}_T^{\mathrm{prox}[i,j]}| \leq g_\infty^{-\ell/2}(T^2\alpha_T)^{-\ell/2}\left(M\frac{(T\pi+1)^\ell}{T} + (1-\alpha_T^{\ell/2})\right) \leq M3^\ell T^{-1},$$

where we used that $T \geq 3$ and $g_\infty\alpha_T \geq 1$, and that $1 - \alpha_T^{\ell/2} = 0$ for $\ell = 0$. Recall the definition (4.19) of $\mathcal{V}_T$ to get $\mathcal{V}_T \leq M3^\ell T^{-1}$. This finishes the proof.

4.7 Proof of Theorem 4.2.3

This section is devoted to the proof of Theorem 4.2.3. Let $T \in \mathbb{N}$ and consider a positive scaling σ_T. In order to prove the theorem we shall apply [Butucea et al., 2022a, Theorem 2.1] replacing the limit kernel, noted $\mathcal{K}_\infty$ therein, by the approximating kernel $\mathcal{K}_T^{\mathrm{prox}}$ defined on Θ^2 by:

$$\mathcal{K}_T^{\mathrm{prox}}(\theta, \theta') = F(|\theta - \theta'|/\sigma_T).$$

We check that all the hypotheses of [Butucea et al., 2022a, Theorem 2.1] hold in our framework. Since Assumption 4.1.1 holds, the noise w_T is admissible and satisfies Point (i) of [Butucea et al., 2022a, Theorem 2.1]. Then, recall that Assumptions 4.2.1 and 4.2.2 are in force thanks to Assumption 4.2.4 (i). Therefore Point (ii) of [Butucea et al., 2022a, Theorem 2.1] on the regularity of the dictionary φ_T is verified. We shall check Point (iii) of [Butucea et al., 2022a, Theorem 2.1] on the regularity of the kernel with $\mathcal{K}_\infty$ replaced by $\mathcal{K}_T^{\mathrm{prox}}$. Since Assumption 4.2.3 holds, we readily check that [Butucea et al., 2022a, Assumption 5.1] as $\mathcal{K}_T^{\mathrm{prox}}(\theta, \theta) = F(0) = 1$ and $\mathcal{K}_T^{\mathrm{prox}[2,0]}(\theta, \theta) = F''(0)/g_\infty = -1$. Thus, Points (iii) therein holds on Θ (with $\Theta_\infty = \Theta$). Point (iv) on the proximity between the kernels $\mathcal{K}_T$ and $\mathcal{K}_T^{\mathrm{prox}}$ is verified since Assumption 4.2.4 (iii) holds and implies [Butucea et al., 2022a, Assumption 5.2].

It remains to show that Point (v) on the existence of certificate functions also holds. To do so, we shall apply [Butucea et al., 2022a, Propositions 7.4 and 7.5] that give sufficient conditions for Point (v) to hold. Let us first focus on the hypotheses of [Butucea et al., 2022a, Proposition 7.4]. We fix $r \in (0, 1/\sqrt{2\, g_\infty\, L_2})$ (we stress that the quantities "r" and "ρ"from [Butucea et al., 2022a, Propositions 7.4 and 7.5] are respectively taken equal to $r\sqrt{g_\infty}$ and 2). It is straightforward to see that Point (i) of [Butucea et al., 2022a, Proposition 7.4] on the regularity of the dictionary is satisfied thanks to Assumption 4.2.2 and 4.2.1.

The Riemannian metric noted $\mathfrak{d}_\infty$ in [Butucea et al., 2022a] is given by, for any $\theta, \theta' \in \Theta$:

$$\mathfrak{d}_\infty(\theta, \theta') = |G_{\mathcal{K}_T^{\mathrm{prox}}}(\theta) - G_{\mathcal{K}_T^{\mathrm{prox}}}(\theta')| = \sqrt{g_{\mathcal{K}_T^{\mathrm{prox}}}}|\theta - \theta'| = \sqrt{g_\infty}\sigma_T^{-1}|\theta - \theta'|, \tag{4.74}$$

where $G_{\mathcal{K}_T^{\mathrm{prox}}}$ is a primitive of the function $\sqrt{g_{\mathcal{K}_T^{\mathrm{prox}}}}$ defined by (4.10) and we used (4.15) for the second inequality. Following [Butucea et al., 2022a, (Eq.39-40)], we define the quantities for $r' > 0$,

$$\varepsilon_\infty(r') = 1 - \sup\left\{|\mathcal{K}_T^{\mathrm{prox}}(\theta, \theta')|; \quad \theta, \theta' \in \Theta \text{ such that } \mathfrak{d}_\infty(\theta', \theta) \geq r'\right\},$$
$$\nu_\infty(r') = -\sup\left\{\mathcal{K}_T^{\mathrm{prox}[0,2]}(\theta, \theta'); \quad \theta, \theta' \in \Theta \text{ such that } \mathfrak{d}_\infty(\theta', \theta) \leq r'\right\}.$$

We readily check that for any $r' > 0$, $\varepsilon(r'/\sqrt{g_\infty}) = \varepsilon_\infty(r')$ and $\nu(r'/\sqrt{g_\infty}) = \nu_\infty(r')$. Thus, $\varepsilon_\infty(r\sqrt{g_\infty}/2) > 0$ and $\nu_\infty(2r\sqrt{g_\infty}) > 0$. Furthermore, Assumption 4.2.3 on the properties of the function F is in force which corresponds to [Butucea et al., 2022a, Assumption 5.1].

Hence, Point (ii) of [Butucea et al., 2022a, Proposition 7.4] on the regularity of the "limit" kernel $\mathcal{K}_T^{\mathrm{prox}}$ holds.

Following [Butucea et al., 2022a, (Eq.42)], we define for $u > 0$:

$$\delta_\infty(u,s) = \inf\Big\{\delta > 0 : \max_{1\le \ell\le s} \sum_{k=1,k\neq\ell}^{s} |\mathcal{K}_T^{\mathrm{prox}[i,j]}(\theta_\ell,\theta_k)| \le u \text{ for all } (i,j)\in\{0,1\}\times\{0,1,2\} \\ \text{and for all } \ell\neq k,\ \mathfrak{d}_\infty(\theta_k,\theta_\ell) > \delta\Big\}.$$

Elementary calculations using (4.14) and (4.74) show that for any $u > 0$, we have $\delta_\infty(u,s) = \sqrt{g_\infty}\,\delta(u,s)$ where δ is defined in (4.20). We fix $u_\infty = \eta H_\infty^{(2)}(r)$. By assumption, we have that $\delta(u_\infty,s) < +\infty$. Therefore, $\delta_\infty(u_\infty,s)$ is finite and Point (iii) of [Butucea et al., 2022a, Proposition 7.4] holds. Recall that we have from Assumption 4.2.4 (iii) that $C_T \le 2$ which gives that Point (iv) of [Butucea et al., 2022a, Proposition 7.4] holds with $\rho = 2$ therein.

We verify Point (v) of [Butucea et al., 2022a, Proposition 7.4] on the proximity between the kernels $\mathcal{K}_T$ and $\mathcal{K}_T^{\mathrm{prox}}$ thanks to Point (iii) of Assumption 4.2.4. We have verified all the assumptions of [Butucea et al., 2022a, Proposition 7.4]. Similarly Points $(i)-(iv)$ of [Butucea et al., 2022a, Proposition 7.5] hold with $u'_\infty = u_\infty$.

Finally, acording to [Butucea et al., 2022a, Propositions 7.4 and 7.5] Point (v) of Theorem [Butucea et al., 2022a, Theorem 2.1] on the existence of certificate functions holds for any subset $\mathcal{Q}^\star$ such that for all $\theta \neq \theta$, we have

$$\mathfrak{d}_T(\theta,\theta) > 2\max(r, 2\delta_\infty(\eta H_\infty^{(2)}(r), s)) = 2\max(r, 2\, g_\infty^{1/2}\delta(\eta H_\infty^{(2)}(r), s)), \tag{4.75}$$

where $\mathfrak{d}_T$ is defined in (4.16).

Recall that by assumption $C_T \le 2$. Since for any $\theta \neq \theta' \in \mathcal{Q}^\star$ we get from the bound (4.18) on $\mathfrak{d}_T$ and Assumption 4.2.4 (iv) that:

$$|\theta-\theta'|/C_T > 4\,\sigma_T\, g_\infty^{-1/2}\ \max(r, 2\, g_\infty^{1/2}\delta(\eta H_\infty^{(2)}(r), s)).$$

Thus, inequality (4.75) holds. We deduce that Point (v) of Theorem [Butucea et al., 2022a, Theorem 2.1] is verified. Finally, by [Butucea et al., 2022a, Theorem 2.1], there exist finite positive constants $\mathcal{C}_0$, $\mathcal{C}_1$, $\mathcal{C}_2'$, $\mathcal{C}_3$, depending on $\mathcal{K}_T^{\mathrm{prox}}$ and on r such that for any $\tau > 0$ and a tuning parameter: $\kappa \ge \mathcal{C}_1\overline{\sigma}\sqrt{\Delta_T \log(\tau)}$, we have the prediction error bound of the estimators $\hat\beta$ and $\hat\vartheta$ defined in (4.23) given by (4.25) with probability larger than $1 - 2\sqrt{g_\infty}\,\mathcal{C}_2'\left(\frac{|\Theta_T|}{\sigma_T\,\tau\sqrt{\log(\tau)}} \vee \frac{1}{\tau}\right)$, where the diameter $|\Theta_T|_{\mathfrak{d}_T}$ of the set Θ_T with respect to the metric $\mathfrak{d}_T$ is bounded by $2\sqrt{g_\infty}|\Theta_T|/\sigma_T$ using (4.18) and the fact that $C_T \le 2$. We set $\mathcal{C}_2 = 2\sqrt{g_\infty}\,\mathcal{C}_2'$. In addition, we have (4.26) with the same probability. A careful reading of the proof of Theorem [Butucea et al., 2022a, Theorem 2.1] shows that the constants $\mathcal{C}_0, \mathcal{C}_1, \mathcal{C}_2', \mathcal{C}_3$ appearing in its statement depend only on the quantities $M_{i,j} = \sup_{\Theta_\infty^2} |\mathcal{K}_T^{\mathrm{prox}[i,j]}|$ with $i,j \in \{0,1,2,3\}$ and on some contants appearing in the properties of the certificates (denoted $C_N, C_N', C_F, C_B, c_N, c_F, c_B$ in [Butucea et al., 2022a]). By [Butucea et al., 2022a, Propositions 7.4 and 7.5], we have that the latter depend only on $\varepsilon_\infty(r\sqrt{g_\infty})$, $\nu_\infty(r\sqrt{g_\infty})$ and $M_{i,j}$ with $i,j \in \{0,1,2,3\}$. We readily show, using (4.14) to see that $M_{i,j}$ depend only on F, that they do not depend on T but on only r and F. This finishes the proof.

5 Modeling infrared spectra : an algorithm for an automatic and simultaneous analysis

Contents

Preamble

Infrared spectroscopy is a widely used technology for nondestructive testing of materials. We propose a novel approach to automatically and simultaneously analyze a dataset of infrared spectra. They are modeled by linear combinations of peaks whose shape and position are parametrized. The observed data consist of linear combinations of the time-discretized peaks with an additive noise. In order to recover the peak parameters, common to all the dataset, and the associated amplitudes, which are specific to each spectrum, we formulate a penalized nonlinear optimization problem. In this context, the penalization ensures that the spectra are recovered using a sparse set of common peaks.
Due to the non-convex nature of the problem and the continuous nature of the parameters, a resolution via standard procedures is out of reach. Therefore, we propose an off-the-grid algorithm with alternating convex optimization updates (to estimate the amplitudes of the peaks) and non-convex steps (to estimate the location and the scale of the peaks). In practice, this gives satisfactory results and provides sparse solutions.
We also study the numerical performances of the algorithm on simulated data and on real infrared spectra. The latter come from polychloroprene rubbers used in a marine environment at different aging levels. Eventually, we use a clustering algorithm in order to identify the peaks corresponding to the chemical components involved in the aging process of this material.

The material for the sections 5.1 to 5.7 of this chapter has been published in [Butucea et al., 2021]. Section 5.8 gives complements on the Frank-Wolfe algorithm and its variants.

5.1 Introduction

Infrared spectra measure the interaction of infrared radiations with the matter. They reveal the presence of chemical substances or functional groups in solid, liquid or gaseous forms. This information on the composition of the matter is essential to prevent failures of materials. Therefore, the use of spectroscopy has become widespread in the industry for nondestructive testing. When a large number of spectra are to be analyzed, an automatic procedure is required. In this chapter, we propose a procedure to automatically identify anomalous aging processes of polychloroprene rubbers in contact with sea water. Principal component analysis or its variants such as the partial least square analysis are often performed on a large dataset of spectra but produce results that are difficult to analyze physically. The spectra have many peaks, each peak corresponding to the absorption of an infrared radiation by a chemical compound. Each peak is characterized by its width, its location and its amplitude. The larger the amplitude of the peak, the more concentrated the chemical compound in the material. Several physical phenomenons imply that these peaks have a shape and a width that depend on the chemical substance ([Hollas, 2004]). When it comes to complex materials, the analysis of a spectrum may require some expertise. It involves determining the location and the width of overlapping peaks that are difficult to distinguish. In this case, the use of a numerical method is necessary. We model the peaks using Gaussian functions (the use of Lorentz functions is also common) and then compute the parameters from the observed spectra. In order to estimate the parameters of the model, curve fitting algorithms are commonly used: it amounts to solve a nonlinear least square problem as in [Aragoni et al., 1995] or [Antonov and Nedeltcheva, 2000]. However, the optimization techniques involved for such ill-conditioned problems require an initialization close to the real values of the parameters of the model. A first guess on the location of peaks and the number of peaks in the model is usually necessary. It would be preferable to have automated procedures allowing to get rid of a prior knowledge of the studied material such as in [Alsmeyer and Marquardt, 2004] or [Kriesten et al., 2008]. In order to give a physical sense to each parameter, it is essential not to over-parametrize the model by adding too many peaks to fit the spectra, particularly in the presence of noise. For this reason, we introduce in this chapter estimators that are solutions of a penalized optimization problem similar to the problem that can be found in [Golbabaee and Poon, 2022]. The penalization favors sparse solutions. Our choice of penalization and the fact that our analysis is run simultaneously on spectra ensures that the spectra are recovered using a sparse set of common peaks.

This work uses recent advances in optimization and statistics to extract physically motivated features from a dataset of infrared spectra observed with an additive noise. Note that our approach does not rely on any prior information on the material being studied. Finally, let us stress that the method presented here can be extended to all peak-shaped models and in particular to numerous branches of spectroscopy.

5.2 Definitions and Notations

Let $d \geq 1$ be the dimension of the space $\Theta \subset \mathbb{R}^d$ of peak parameters. The set Θ is assumed to be a compact convex set. Let φ be a positive smooth function defined on $\Theta \times \mathbb{R}$ modeling the shape of peaks. We denote by $T \in \mathbb{N}$ the size of the discretization grid over a wavenumber interval. For a discretization scheme on the real line $(\sigma_j)_{0 \leq j \leq T}$, we set $\Delta_T = (\sigma_T - \sigma_0)/T$. For f, g measurable functions defined on $\mathbb{R}$, we set:

$$\langle f, g \rangle_T = \Delta_T \sum_{j=1}^{T} f(\sigma_j) g(\sigma_j),$$

and also $\|f\|_T = \langle f, f\rangle_T^{1/2}$. We assume that $\|\varphi(\theta)\|_T$ is non zero for all $T \in \mathbb{N}$ and $\theta \in \Theta$. We also define the normalized function ϕ_T on Θ taking values in $\mathbb{R}^T$ by:

$$\phi_T(\theta) = \|\varphi(\theta)\|_T^{-1} \left(\varphi(\theta, \sigma_1), \cdots, \varphi(\theta, \sigma_T)\right).$$

Let $K \in \mathbb{N}^*$. For $\vartheta = (\theta_1, \ldots, \theta_K) \in \Theta^K$, we define the function Φ_T on Θ^K taking its values in $\mathbb{R}^{K\times T}$ by $\Phi_T(\vartheta) = \left(\phi_T(\theta_1)^\top, \cdots, \phi_T(\theta_K)^\top\right)^\top$. When there is no risk of confusion, we write Φ and ϕ instead of Φ_T and ϕ_T, respectively.

5.3 The Model

In this chapter, we model infrared spectra using linear combinations of parametric functions φ (called peaks) with an additive noise. The parametric functions can be Gaussian, Lorentz or Voigt functions, as usually done in the literature. These functions are quite popular for curve fitting or peak deconvolution in many branches of spectroscopy. We refer to [Jansson, 1984] and [Hollas, 2004] for discussions on the physical effects justifying such spectral line shapes (also referred to as spectral line profiles). The location and the width of a peak are specific to a chemical group whereas its amplitude which encodes the concentration of the group depends on the material. Here we lead our study with Gaussian peaks:

$$\begin{aligned} \varphi_G \colon \Theta \subset \mathbb{R}^2 &\to L^2(\mathbb{R}) \\ (\mu, \nu) &\mapsto \mathrm{e}^{-\frac{(\cdot-\mu)^2}{2\nu^2}}. \end{aligned} \tag{Gauss}$$

Other models can be found in the literature such as Lorenztian peaks:

$$\begin{aligned} \varphi_L \colon \Theta \subset \mathbb{R}^2 &\to L^2(\mathbb{R}) \\ (\mu, \nu) &\mapsto \frac{1}{1 + \left(\frac{\cdot-\mu}{\nu}\right)^2}, \end{aligned} \tag{Lorentz}$$

or Voigt peaks obtained from the convolution of a Gaussian and a Lorentzian peak:

$$\begin{aligned} \varphi_V \colon \Theta \subset \mathbb{R}^3 &\to L^2(\mathbb{R}) \\ (\mu, \nu_1, \nu_2) &\mapsto \int_{-\infty}^{+\infty} \frac{\mathrm{e}^{-\frac{t^2}{2\nu_1^2}}}{1 + \left(\frac{\cdot-t-\mu}{\nu_2}\right)^2} \,\mathrm{d}t. \end{aligned} \tag{Voigt}$$

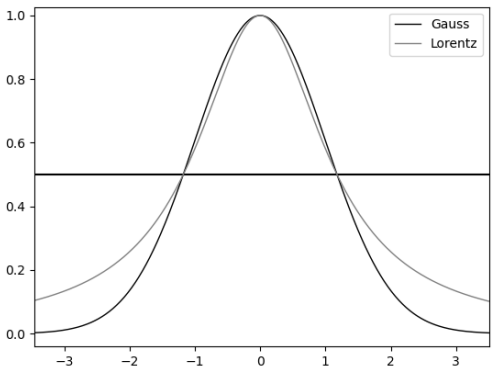

Figure 5.1 – Gaussian and Lorentzian profiles used to model absoption peaks represented with the same half-width (Gaussian peak with $\mu = 0, \nu = 1$ and Lorentzian peak with $\mu = 0, \nu = \sqrt{2\log(2)}$).

Consider a data set of n spectra $(y_i)_{1\leq i\leq n}$ disctretized on T wavenumbers $(\sigma_j)_{1\leq j\leq T}$, then write for all $1\leq i\leq n, 1\leq j\leq T$:

$$y_i(\sigma_j)=\sum_{k=1}^{K} B^\star_{i,k}\frac{\varphi(\theta^\star_k,\sigma_j)}{\|\varphi(\theta^\star_k)\|_T}+W_{ij}. \tag{5.1}$$

In model (5.1), $(W_{ij})_{1\leq i\leq n,1\leq j\leq T}$ denote the random variables modeling the noise, assumed to be independent Gaussian variables with zero-mean and variance $\overline{\sigma}^2$. The row vectors $(B^\star_{i,.})_{1\leq i\leq n}\in\mathbb{R}^K_+$ of the matrix $B^\star$ have their k^{th}coordinate encoding the amplitude of the k^{th} peak involved in the linear combination. The positivity of their entries is physically motivated by the fact that spectra can only take positive values. The peaks are shared by all the spectra in the dataset but their amplitudes are specific to each spectrum. Note that each spectrum individually may have only a few peaks with non zero amplitudes. We denote by K an upper bound of the number of peaks which can be arbitrarily large. Since the family of Gaussian functions $(\varphi(\theta),\ \theta\in\Theta)$ is linearly independent, the spectra decomposition is unique.

The model (5.1) can be written in matrix form:

$$Y=B^\star\Phi(\vartheta^\star)+W,$$

where $Y\in\mathbb{R}^{n\times T}, Y_{ij}=y_i(\sigma_j)$, $W\in\mathbb{R}^{n\times T}$, $B^\star\in\mathbb{R}^{n\times K}_+$, $\vartheta^\star=(\theta^\star_1,\dots,\theta^\star_K)$. One can notice that applying the same permutation on the columns of $B^\star$ and the coordinates of $\vartheta^\star$ gives the same model. It amounts to change the order of the peaks in the linear combinations of (5.1). The matrix $B^\star$ and the K-uplet $\vartheta^\star$ are defined up to such a joint permutation.

The spectra are expected to be decomposed in a small number of active peaks, that is why the matrix $B^\star$ is sparse and have numerous zero entries. Moreover, they are expected to have similarities so that only a few peaks are used to model the whole dataset. Hence, the matrix $B^\star$ have many columns set to zero. We denote by $S^\star$ the indices of the non zero columns of $B^\star$ which feature the active peaks in the dataset:

$$S^\star=\{k,\ \text{there exists } 1\leq i\leq n, B^\star_{ik}\neq 0\}.$$

Thus, the peaks $\varphi(\theta^\star_k)$ whose index k does not belong to the set $S^\star$ play no role in the model. We denote by $\vartheta^\star_{S^\star}$ the restriction of $\vartheta^\star$ to coordinates whose indices belong to $S^\star$.

5.4 Optimization Problem

In this section, we retrieve the nonlinear parameters $\vartheta^\star_{S^\star}$ (encoding the active peak), as well as the linear parameters $B^\star$ (encoding the amplitudes of peaks) that fully describe the model. We formulate a program similar to the group-Lasso problem for sparse linear models introduced in [Yuan and Lin, 2006] and discussed in many papers since then ([Lounici et al., 2011], [Obozinski et al., 2011]). We generalize it to a larger range of sparse models in the same way as [Boyer et al., 2017] and more recently as [Golbabaee and Poon, 2022]. This leads to a nonlinear least square problem with a penalization term weighted by a real parameter $\lambda>0$:

$$\min_{\substack{B\in\mathbb{R}^{n\times K}_+\\ \vartheta\in\Theta_{K,T}(h)}}\mathcal{F}_{\lambda,\varphi}(B,\vartheta),\quad\text{where}\quad\mathcal{F}_{\lambda,\varphi}(B,\vartheta):=\frac{1}{nT}\|Y-B\Phi(\vartheta)\|^2_{\ell_2}+\lambda\|B\|_{\ell_1,\ell_2} \tag{5.2}$$

and

- $\|\cdot\|_{\ell_2}$ is the usual Euclidean norm and $\|B\|_{\ell_1,\ell_2}=\sum_{k=1}^{K}\|B_{\cdot,k}\|_{\ell_2}$ is the mixed (1,2)-norm,

- $\Theta_{K,T}(h) \subset \Theta^K$ with $h > 0$, is the set of parameters $\vartheta = (\theta_1, \cdots, \theta_K) \in \Theta^K$ such that for all $1 \leq \ell, k \leq K, \ell \neq k$:

$$\mathcal{K}_T(\theta_\ell, \theta_k) := \frac{|\langle \varphi(\theta_\ell), \varphi(\theta_k) \rangle_T|}{\|\varphi(\theta_\ell)\|_T \, \|\varphi(\theta_k)\|_T} < h.$$

Let $(\hat{B}(\lambda), \hat{\vartheta}(\lambda))$ be solution of the problem (5.2) (or simply $(\hat{B}, \hat{\vartheta})$ when there is no ambiguity). We denote by $\hat{\vartheta}_{\hat{S}}$ the restriction of $\hat{\vartheta}$ to coordinates whose indices belong to:

$$\hat{S} = \{k : \text{ there exists } 1 \leq i \leq n, \hat{B}_{ik} \neq 0\}.$$

The set $\hat{S}$ gathers the indices of the active peaks used to fit the spectra. The penalization used in model (5.2) promotes group sparsity in the sense that it favors a matrix B that has columns with zero entries. This leads to solutions that use fewer peaks while correctly approximating the data. We refer to [Obozinski et al., 2011] and [Lounici et al., 2011] to better understand the penalization procedure in the case where the parameters $\vartheta^\star$ are known (but $S^\star$ is unknown). We also enforce the positivity of the linear parameters with constraints on the entries of the matrix $\hat{B}$.

The set $\Theta_{K,T}(h)$ introduced in this chapter corresponds to a separation criterion on the peaks used to fit the data, the separation being measured by h. Provided that h is small enough, the matrix $\Phi(\hat{\vartheta})\Phi(\hat{\vartheta})^\top$ is of full rank. This is for example the case if $h < 1/(K-1)$, thanks to Gershgorin's theorem. This implies then that $\hat{B}$ is the unique solution of the problem:

$$\min_{B \in \mathbb{R}_+^{n \times K}} \frac{1}{nT} \left\| Y - B\Phi(\hat{\vartheta}) \right\|_{\ell_2}^2 + \lambda \|B\|_{\ell_1,\ell_2},$$

which amounts to minimize a strictly convex function over a convex set. The value chosen for h is a compromise: it must be large enough to estimate overlapping peaks in the dataset but sufficiently small to make the model identifiable in terms of the linear parameters. The identifiability for linear parameters may be crucial to give them a physical sense.

5.5 Algorithm

5.5.1 Presentation of the algorithm

The resolution of the optimization problem (5.2) is not an easy task at first glance since the optimization problem is non-convex. Indeed, the peaks are not even convex in terms of their parameters. A brutal gradient-descent would be hopeless without a very good initialization. In this chapter, we want to proceed without any prior knowledge on the parameters to be estimated. It might be tempting to use a grid on the space of nonlinear parameters describing the peaks and use sparse methods to retrieve the amplitudes as suggested in [Tang et al., 2013a]. But, the approximation of the spectra would depend on the chosen grid. Typically, it would be impossible to recover exactly the parameters of a peak without an infinitely thin grid. That is why an off-the-grid algorithm which does not discretize the parameter space must be preferred. Thus, it will be possible to recover exactly the parameters of the peaks provided their overlap is low (characterized by the parameter h). Recent progress in optimization have shown the efficiency of off-the-grid procedures such as the Sliding Frank-Wolfe iterations (see [Denoyelle et al., 2020]) or the alternating descent conditional gradient method (see [Boyd et al., 2017]). Both algorithms are based on the addition of a new peak at each iteration to approximate one spectrum. During an iteration, a new peak is placed, then all the parameters are re-estimated with an improved initialization. The work of [Golbabaee and Poon, 2022] extended the Frank-Wolfe algorithm to fit several spectra simultaneously. We propose a variant of the Sliding Frank-Wolfe algorithm, see Algorithm 1 below, that separates the optimization of linear and nonlinear parameters and which merges peaks that

are highly overlapping. This allows the use of classical algorithms to solve a standard group-Lasso problem for linear parameters. Hence, this approach takes advantage of the fact that linear parameters are often more numerous than non linear parameters ($n \times K$ v.s $d \times K$) and always much easier to compute.

Algorithm 1: Sliding Frank-Wolfe iterations

Data: Y
Input: φ, λ, h
Output: ϑ, B
Initialize: $i := 0$, $R^{(0)} := Y$, $\vartheta^{(0)} := \emptyset$
while $i < K$ **do**

$\theta^{(i+\frac{1}{2})} \in \underset{\theta \in \Theta}{\operatorname{argmax}} \left\| R^{(i)} \phi(\theta)^\top \right\|_{\ell_2}^2$

$\vartheta^{(i+\frac{1}{2})} = \left(\vartheta^{(i)}, \theta^{(i+\frac{1}{2})} \right)$ `// Adding new peak`

$B^{(i+\frac{1}{2})} \in \underset{B \in \mathbb{R}_+^{n \times (i+1)}}{\operatorname{argmin}} \mathcal{F}_{\lambda,\varphi}(B, \vartheta^{(i+\frac{1}{2})})$ `// Convex step`

$\vartheta^{(i+1)} \in \underset{\vartheta \in \Theta^{i+1}}{\operatorname{argmin}} \mathcal{F}_{\lambda,\varphi}(B^{(i+\frac{1}{2})}, \vartheta)$ initialized in $\vartheta^{(i+\frac{1}{2})}$ `// Non-convex step`

Merging routine $(\vartheta^{(i+1)}, h)$ `// Merging overlapping peaks and adding peaks with parameters chosen at random`

$B^{(i+1)} \in \underset{B \in \mathbb{R}_+^{n \times (i+1)}}{\operatorname{argmin}} \mathcal{F}_{\lambda,\varphi}(B, \vartheta^{(i+1)})$ `// Re-estimation of linear parameters`

$R^{(i+1)} = Y - B^{(i+1)} \Phi(\vartheta^{(i+1)})$
$i = i + 1$
end

Algorithm 2: A merging routine

Input: $(\theta_1, \ldots, \theta_m)$, h
Output: $(\theta_1, \ldots, \theta_m)$
while $(\theta_1, \ldots, \theta_m) \notin \Theta_{m,T}(h)$ **do**
  for $1 \leq \ell < k \leq m$ **do**
    if $\mathcal{K}_T(\theta_\ell, \theta_k) > h$ **then**
      θ_k is chosen at random in the parameter space
    end
  end
end

Remark 5.5.1. In the Sliding Frank-Wolfe algorithm as introduced in [Denoyelle et al., 2020], the peaks are never merged and therefore the re-estimation of linear and nonlinear parameters in Algorithm 1 can be done simultaneously. Splitting into two steps avoids a gradient descent on all the parameters simultaneously.

5.5.2 Implementation details

We used a L-BFGS-B algorithm for the non-convex steps, see [Byrd et al., 1995] and [Zhu et al., 1997]. The method does not require the knowledge of the Hessian of the objective function. It is a quasi-Newton method that uses an approximation of the Hessian. We refer to [Nocedal and Wright, 2006] for a review on quasi-Newton methods. The algorithm

allows the addition of constraints. Typically, peaks that are too thin to appear between two discretization points or wide peaks covering the whole range of observation should not be taken into account in the optimization. In particular, when one uses the Gaussian function $\varphi_G(\mu,\nu) = \mathrm{e}^{-(\cdot-\mu)^2/(2\nu^2)}$ to model the peaks, one can take $2\nu > \Delta_T$ so that a peak has a significant contribution on the discretization points. As for an upper bound on ν, one can take for the Gaussian model $6\nu < \sigma_T - \sigma_1$ so that a Gaussian function at the center of the observation range puts at least 99% of its mass between σ_1 and σ_T. It is also legitimate to require that the location parameter μ belongs to the range of observations *i.e.*: $\sigma_1 \leq \mu \leq \sigma_T$.

Without any prior information on the overlapping of the peaks, it may be necessary to re-run the algorithm and decrease h until one gets $\Phi(\hat{\vartheta})$ of full rank.

5.6 Numerical Applications

5.6.1 Simulated data

We tested the Algorithm 1 on noisy spectra composed of at most 15 Gaussian peaks. We generated a set of $n = 10$ spectra within the range $\sigma_{\min} = 0$ to $\sigma_{\max} = 20$. The parameters location μ and scale ν of the 15 Gaussian peaks were chosen at random according to a uniform distribution on the parameter space ($5 \leq \mu \leq 15$ and $10 \cdot \Delta_T \leq \nu \leq \frac{T \cdot \Delta_T}{6}$). We considered a Gaussian noise on the spectra by adding at each point of the discretization independent and identically distributed Gaussian random variables of mean 0 and standard deviation $\overline{\sigma} \in \{0, 0.01, 0.1, 0.5\}$ (see Figure 5.2). In order to show the consistency of the method, we computed for the different values of $\overline{\sigma}$, the mean squared error between the data reconstructed with the estimated parameters and the data without noise,

$$MSE^* = \frac{1}{nT}\|B^\star\Phi(\vartheta^\star) - \hat{B}\Phi(\hat{\vartheta})\|_{\ell_2}^2.$$

The estimated parameter $(\hat{B}, \hat{\vartheta})$ depends on the penalization parameter λ taken in the optimization problem (5.2). We took for λ the orders of magnitude that lead to the optimal convergence rates in the case of linear models as shown in [Lounici et al., 2011] ($\lambda \sim \overline{\sigma}/\sqrt{nT}$) and we took $\lambda = 0.01/\sqrt{nT}$ for $\overline{\sigma} = 0$. The values of MSE^* from Figure 5.3 show that we managed to reconstruct almost exaclty the spectra in less than 100 iterations of the algorithm when $\overline{\sigma} \leq 0.01$.

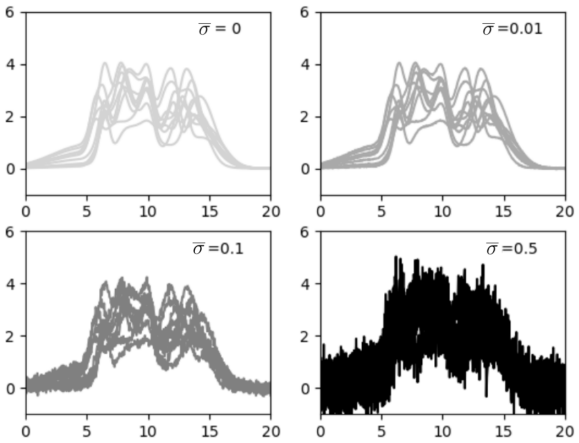

Figure 5.2 – Representation of the simulated data with different noise levels ($\overline{\sigma} \in \{0, 0.01, 01, 0.5\}$).

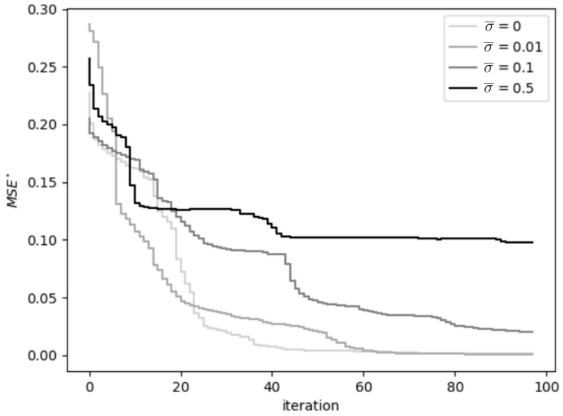

Figure 5.3 – Evolution of MSE^* over the iterations of the algorithm for different noise levels.

5.6.2 Aging of polychloroprene rubbers

5.6.2.1 Presentation of the dataset

The data used in our study were obtained from spectroscopic analysis of samples of polychloroprene rubbers, one side of which was in contact with seawater and the other was glued to steel. The device to obtain the spectra is a Fourier-transform infrared spectrometer in Attenuated Total Reflectance (ATR) mode. The spectra, visualized in a graph of infrared light absorbance on the vertical axis vs. wavenumbers on the horizontal axis, have to be normalized for the quantitative analysis of peak amplitudes. In addition, a pre-processing is also performed to remove the baselines present on the spectra. The multiplicative normalization is specific to each spectra and such that its peak amplitude for the $C-Cl$ bond situated at 825 cm^{-1} is equal to 1. The $C-Cl$ bond was chosen for the normalization of the spectra because of its stability with respect to the aging process, see [Le Gac et al., 2012]. It is then possible to compare the peak amplitudes between the spectra in the dataset (see Figure 5.4).

The 72 spectra composing the dataset are discretized between 4000 cm^{-1} and 600 cm^{-1}

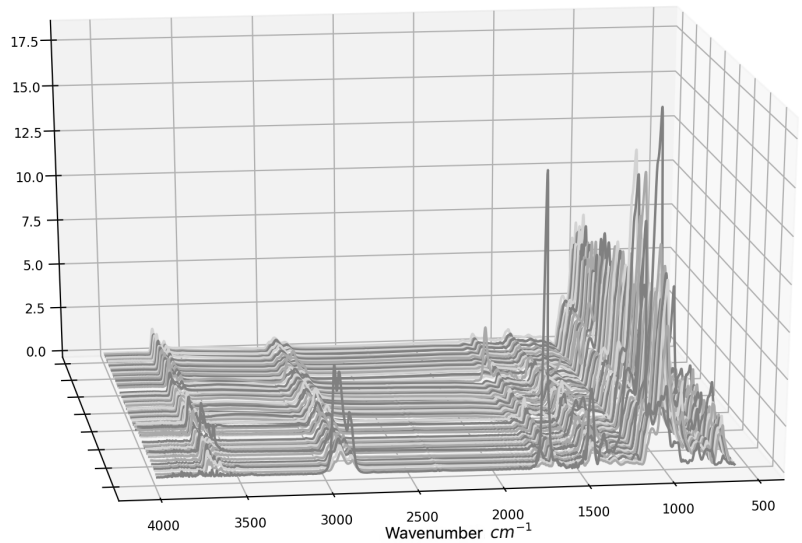

Figure 5.4 – Representation of all the infrared spectra of polychloroprene samples after normalization and removal of baselines.

with measures every 2 cm^{-1}. We focus our analysis on the area between 2000 cm^{-1} and 600 cm^{-1} where are the biggest dissimilarities between the spectra.

Table 5.1 – Table of the locations of peaks and their corresponding bonds for the polychloroprene samples. The values are taken from [Tchalla, 2017].

Wavenumbers (cm^{-1})	Peak assignment
3690-3400-3364 3200-3014	-OH
2952-2920-2850	$\nu - CH_2, CH_3$ Aliphatic
1731	$\nu - C = O$
1647	$\nu - C = C$ of $HC = CH_2$
1540	$\nu - C = C$ of $R - CR = CH - R$ and $\delta - CH_2$ Aliphatic
1419	$\delta - CH_2, \delta - CH$ Aliphatic
1160-1082	$\nu - Si - O$ (SiO_2)
1009-909	$\nu - Si - O$ $(Si - OH)$
825	$C - Cl$
664	CH Aromatic

5.6.2.2 Aging properties of polychloroprene rubbers

Polychroroprene is often used in marine structure to prevent corrosion. Some works on polychloroprene rubbers in marine environments have brought out some physical phenomenons that occur with aging (see [Le Gac et al., 2012], [Tchalla et al., 2017]). The sea water diffuses into the material until it is saturated. During the process, several reactions might appear. In [Le Gac et al., 2012], the authors have pointed out the hydrolysis of silica fillers. It consists in a formation of silanol from the silica fillers. This reaction is reflected in the spectra by a decrease of the $1160 - 1082$ peaks (attributed to the $Si - O$ bond of the silica filler) and a new peak located around 1009 cm^{-1}(attributed to $Si - OH$) that rises according to the aging duration. In addition, they showed that a carbonyl formation can occur due to an oxidation reaction. This can be seen in the spectra by the appearance of a new peak at 1731 cm^{-1}. In [Tchalla, 2017], peaks in spectra from polychloroprene samples with aging conditions similar to those in our study are attributed to their corresponding chemical bond (see Table 5.1).

5.6.2.3 Estimation of linear and nonlinear parameters for the peak-shaped model

In the optimization problem (5.2) of Section 5.4, the penalization parameter λ must be tuned. Intuitively, choosing a large value for λ will make the term of penalization in (5.2) preponderant and set a lot of entries of the matrix $\hat{B}$ to zero. In this case, one expects the solutions to underestimate the number of peaks in the model. On the contrary, a small value for λ will set very few entries of the matrix $\hat{B}$ to zero and will lead to overestimate the number of peaks in the model. There is no easy way to choose the penalization parameter λ. To achieve a compromise between the number of peaks used and the quality of the spectra approximation,we ran the algorithm on the set of polychloroprene spectra for different values of the tuning parameter λ. It appeared that for the Gaussian model, around $\lambda \approx 5 \cdot 10^{-5}$, the unpenalized mean squared error $\mathcal{F}_{0,\varphi_G}(\hat{B}(\lambda), \hat{\vartheta}(\lambda))$ as well as the penalized mean squared error $\mathcal{F}_{\lambda,\varphi_G}(\hat{B}(\lambda), \hat{\vartheta}(\lambda))$ increase drastically (see Figure 5.5). From this point, the number of peaks used to fit the data drops (see Figure 5.6). Hence, a reasonable choice for λ, is under this critical point. We ran the Algorithm 1 with the Gaussian model ($\varphi := \varphi_G$) for $\lambda = 3 \cdot 10^{-5}$ and $h = 0.9$. We imposed that the location parameter belongs to the range of obervations $[600cm^{-1}, 2000cm^{-1}]$, and that the width parameter ν belongs to $[\frac{\Delta_T}{2} \approx 1, \frac{T\Delta_T}{6} \approx 233]$. Finally, we obtained 66 active peaks for the whole dataset in the range $[600cm^{-1}, 2000cm^{-1}]$ after 100 iterations of the algorithm.

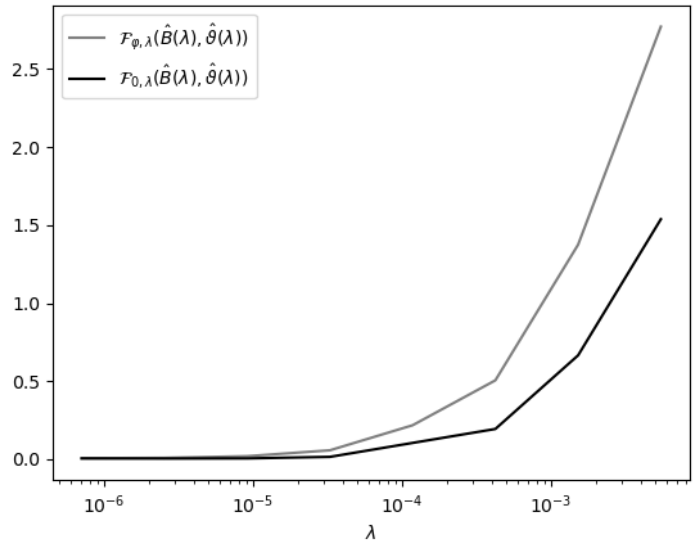

Figure 5.5 – Mean squared error $\mathcal{F}_{0,\varphi}(\hat{B}(\lambda),\hat{\vartheta}(\lambda))$ and penalized mean squared error $\mathcal{F}_{\lambda,\varphi}(\hat{B}(\lambda),\hat{\vartheta}(\lambda))$ seen as functions of the tuning parameter λ.

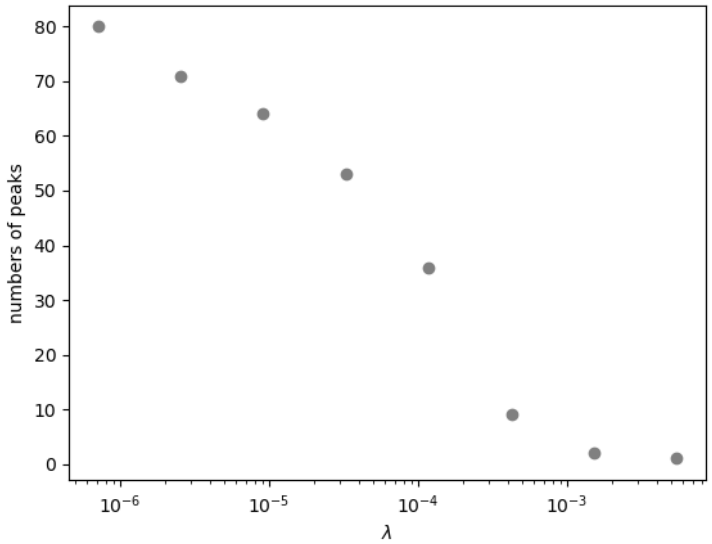

Figure 5.6 – Number of peaks found by the algorithm to fit the spectra of polychloroprene samples as a function of the tuning parameter λ.

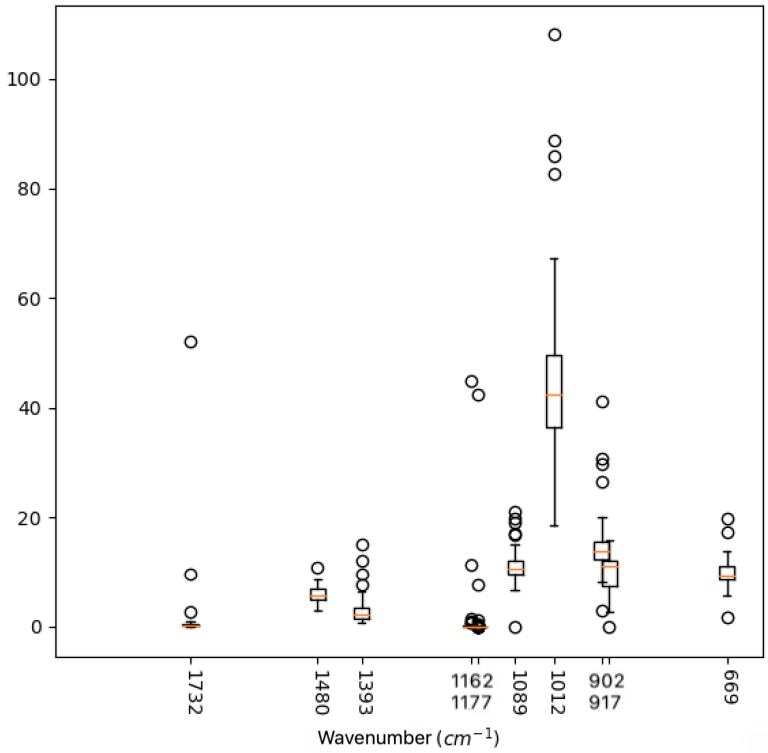

Figure 5.7 – Boxplot for the amplitudes of the 10 most significant peaks for the 72 polychloroprene spectra in the dataset.

5.6.2.4 Boxplots for main peak amplitudes

To understand the distribution of the peak amplitudes within the dataset, we represented them on Figure 5.7 with boxplots. The locations of peaks on the x-axis as well as the amplitude values in the boxes, correspond to those estimated by Algorithm 1. For the sake of readability, we have represented only the ten most significant peaks (those with the biggest sum of squared amplitudes). First, one can notice that Algorithm 1 retrieves peaks close to those referenced in Table 5.1: the carbonyl peaks at 1732 cm^{-1} (vs 1731 cm^{-1} in the Table 5.1) as well as the silica peaks at $1162 - 1089$ cm^{-1} (versus $1160 - 1082$ cm^{-1} in Table 5.1) and the silanol peaks at $1012 - 917 - 902$ cm^{-1} (versus $1009 - 909$ cm^{-1} in Table 5.1). Secondly, it appears that the silanol peak brings the most dissimilarity among the spectra.

5.6.2.5 Clustering on peak amplitudes

In order to bring out different levels of aging among the spectra, a clustering algorithm such as a k-means algorithm whose inputs are the vectors of estimated peak amplitudes $\hat{B}_{i,.} \in \mathbb{R}^K, 1 \leq i \leq n$ can be used. The k-means algorithm aims to partition the n observation vectors into M sets $\mathcal{A} = \{\mathcal{A}_1, \cdots, \mathcal{A}_M\}$ so as to minimize the within-cluster sum of squares. It amounts to solve

$$\min_{\mathcal{A}} \sum_{\ell=1}^{M} \sum_{i \in A_\ell} \left\| \hat{B}_{i,.} - \beta_\ell \right\|_{\ell_2}^2$$

where the vectors β_ℓ are the centroids of the sets $(\mathcal{A}_\ell)_{1 \leq l \leq M}$. Let us write $(\hat{\mathcal{A}}_1, \cdots, \hat{\mathcal{A}}_M)$ the partition returned by a k-means algorithm and $(\hat{\beta}_1, ...\hat{\beta}_M)$ the associated centroids.

The number of clusters is an input of the algorithm. Therefore, one of the first issue to address is related to the number of clusters used. A compromise must be found between gathering the data in a few groups and not having too much dissimilarity within the group. To tackle this issue, solving the k-means problem for different values of M can be useful. We

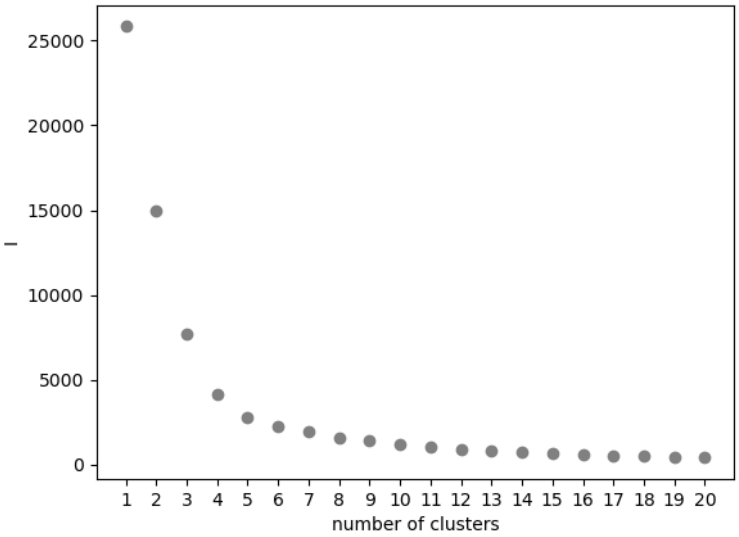

Figure 5.8 – Sum of squared distances of observations to their closest cluster centroid with respect to the numbers of clusters.

plotted in Figure 5.8 the value I(M), as a function of M, of the sum of squared distances of the observations to their closest cluster centroid:

$$I(M) = \sum_{\ell=1}^{M} \sum_{i \in \hat{A}_\ell} \left\| \hat{B}_{i,\cdot} - \hat{\beta}_\ell \right\|_{\ell_2}^2 .$$

By taking $M = 4$ where the curve makes an elbow, we separate the data into a number of clusters small enough to be informative while having drastically reduced the sum of squared distances between observations and their associated centroid (see Figure 5.9). Let us observe

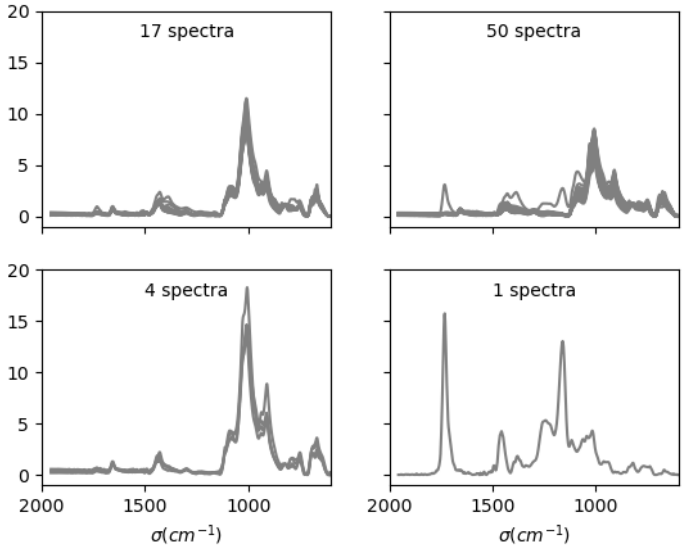

Figure 5.9 – Representation of the polychloroprene spectra within their cluster after running a k-means algorithm on the row vectors $\hat{B}_{i,\cdot} \in \mathbb{R}^K, 1 \leq i \leq n$.

that among the four clusters, one gathers about 70% of the data (top right hand graphic in Figure 5.9). One can also notice that the clusters at the top in Figure 5.9), gathering more

than 90%, are characterized by lower amplitudes for the silanol peak at 1009 cm^{-1}. The spectrum that forms a single-point-cluster presents strong carbonyl peaks centered around 1731 cm^{-1}. Hence, we managed to isolate a spectra that presents a really high level of carbonyl and separate the others with respect to the amplitudes of the silanol and silica peaks without any prior information on the material. Let us recall that the rise of the silanol and carbonyl peaks correspond to reactions involved in the aging process. We also considered the clusters based on the amplitudes of the peaks corresponding only to silica, silanol and carbonyl by selecting the peaks returned by the algorithm that are located at less than 10 cm^{-1} of the positions referenced in the Table 5.1. The clusters obtained correspond exactly to those from Figure 5.9. Therefore, one can conclude that the main differences between the spectra are due to the peaks of carbonyl, silanol and silica which were identified in the literature as involved in the aging process of polychloroprene rubbers in a marine environment. The two clusters at the bottom of Figure 5.9 gather spectra with chemical characteristics of higher aging levels and represent less than 10% of the data.

5.7 Conclusion

In this chapter, we estimate an arbitrary number of infrared spectra simultaneously without any prior information. The spectra are modeled under the physical constraints by linear combinations of peaks and each peak belongs to a nonlinear parametric family of functions (e.g. Gaussian). The estimation consists in a generalization to nonlinear models of the group-Lasso optimization problem. This formulation allows to limit the number of peaks used to fit the data. A numerical method is proposed with an off-the-grid scheme. The limited resolution problems, intrinsic to the use of a grid on the parameter space, are thus avoided. The method is numerically consistent in the presence of noise and favors sparse solutions. Although the problem is nonlinear, the method works without any special care for the initialization. Moreover, the computation time behaves well with a large number of spectra as long as the number of peaks to fit the data does not increase drastically. We apply this approach to real data of polychloroprene rubber spectra, and recover the main peaks associated with its chemical components and identify by clustering those involved in its aging process. The locations of the peaks found by the algorithm are consistent with those established by previous work in the field of chemistry.

5.8 Complements on the Frank-Wolfe algorithm and its variants

In this section we shall focus on the resolution of the problem (5.2). To be more general, we shall consider the optimization methods performed over the set of matrices having real entries. Therefore we shall rather consider for a real tuning parameter λ the problem:

$$\min_{\substack{B \in \mathbb{R}^{n \times K} \\ \vartheta \in \Theta_{K,T}(h)}} \mathcal{F}_{\lambda,\varphi}(B, \vartheta^{\mathcal{G}}), \tag{5.3}$$

where $\mathcal{F}_{\lambda,\varphi}(B, \vartheta^{\mathcal{G}})$ is defined in (5.2). We shall stress when the methods adapt to the case where the minimization is performed over matrices having nonnegative entries as in (5.2).

5.8.1 Towards off-the-grid-algorithms

In order to solve (5.3), one could use a grid on the space of nonlinear parameters describing the peaks and then use sparse methods to retrieve the amplitudes, as proposed in [Tang et al., 2013a]. Indeed, it would amount to solve the problem:

$$\min_{B \in \mathbb{R}^{n \times m}} \mathcal{F}_{\lambda,\varphi}(B, \vartheta^{\mathcal{G}}), \tag{5.4}$$

where $\vartheta^{\mathcal{G}} = (\theta_1^{\mathcal{G}}, \cdots, \theta_m^{\mathcal{G}})$ with $(\theta_i^{\mathcal{G}}, 1 \leq i \leq m)$ a given grid on the parameter space Θ. The problem reduces to a linear optimization problem penalized by a mixed norm on the matrix of linear coefficients. There are efficient numerical methods for solving the general case (5.4) as well as the constrained case where the linear coefficients are required to be positive. We mention the fast iterative shrinkage-thresholding algorithms (FISTA, see [Beck and Teboulle, 2009]) that are modifications of the previous iterative shrinkage-thresholding algorithms (ISTA). The latter consists in a classical gradient descent where a soft thresholding operator is applied at each iteration on the current approximation of the minimum. It produces a sequence $(B^{(k)}, 1 \leq k \leq K)$ approximating a minimum $B^\star$ of the functional $\mathcal{F}_{\lambda,\varphi}(\cdot, \vartheta^{\mathcal{G}})$ satisfying:

$$\mathcal{F}_{\lambda,\varphi}(B^{(k)}, \vartheta^{\mathcal{G}}) - \mathcal{F}_{\lambda,\varphi}(B^\star, \vartheta^{\mathcal{G}}) = \mathcal{O}(1/k).$$

In the improved FISTA version, the thresholding operator used to obtain $B^{(k)}$ is not applied to $B^{(k-1)}$ but rather to a well chosen linear combination of $B^{(k-1)}$ and $B^{(k-2)}$. This method reaches the better global convergence rate

$$\mathcal{F}_{\lambda,\varphi}(B^{(k)}, \vartheta^{\mathcal{G}}) - \mathcal{F}_{\lambda,\varphi}(B^\star, \vartheta^{\mathcal{G}}) = \mathcal{O}(1/k^2).$$

We refer to [Beck and Teboulle, 2009] for a proof of these results in a more general framework.

Nevertheless, these methods suffer from the discretization of the parameter space. The solution to the problem (5.3) may suffer a large additionnal bias if the nonlinear parameters do not belong to the grid. A natural idea to improve the approximation of a solution of (5.3) is to increase the number of points in the grid, *i.e* the dimension of $\vartheta^{\mathcal{G}}$. However, it does not come without a cost. Typically, in the case of ISTA and FISTA algorithms, the rate of convergence is ruled by a multiplicative constant that is bounded from below by the largest eigenvalue of $\Phi(\vartheta^{\mathcal{G}})\Phi(\vartheta^{\mathcal{G}})^\top$, see Theorem 4.4 and Example 2.2 of [Beck and Teboulle, 2009] in this direction. Thus increasing the number of points in the grid increases the correlation between the rows of the matrix $\Phi(\vartheta^{\mathcal{G}})$ and deteriorates the rate of convergence of the algorithm. Furthermore, increasing the correlation between the rows of $\Phi(\vartheta^{\mathcal{G}})$ may lead to a violation of the constraint $\vartheta \in \Theta_{K,T}(h)$ from (5.3) implying that the peaks used to fit the spectra have a pairwise correlation inferior to h. We also remark that when the number of points in the grid m is larger than the bound K on the true number of active peaks $s = \mathrm{Card}(S^\star)$, there is no guarantee to retrieve a solution with less than K peaks. In [Duval and Peyré, 2017a], another major drawback of the discretized problem (5.4) is pointed out. Indeed the authors show that the method tends to produce clusters of peaks around the true peaks one seek to estimate (see Theorem 2 and Corollary 1 therein). Let us also mention that when the dimension d of the parameter space Θ increases, the number of points needed in the grid increases drastically. It may lead to numerical issues even if the problem to solve (5.4) is linear. For all these reasons, we shall consider methods that do not rely on a grid on the parameter space Θ. We shall refer to these methods as "off-the-grid" methods.

5.8.2 The Frank-Wolfe algorithm and its variants

In this section, we shall present a method that solves efficiently (5.3) in a gridless manner. We shall first consider the case where the peaks returned by the procedure are not required to have a small pairwise correlation and when only one spectrum is to be estimated, *i.e* $h > 1$ (so that $\Theta_{K,T}(h) = \Theta^K$) and $n = 1$. In this case, we can consider the more general optimization problem introduced in [de Castro and Gamboa, 2012] and refered to as the Beurling Lasso (or BLasso). It writes for a real positive tuning parameter λ:

$$\min_{\mu \in \mathcal{M}(\Theta)} \quad \mathcal{G}_{\lambda,\varphi}(\mu), \qquad \text{where} \qquad \mathcal{G}_{\lambda,\varphi}(\mu) := \frac{1}{T}\|Y - \Phi\mu\|_{\ell_2}^2 + \lambda\|\mu\|_{TV}, \tag{5.5}$$

and:

- $\|\cdot\|_{TV}$ is the total variation norm on measures,
- $\mathcal{M}(\Theta)$ denotes the set of signed Radon measures on the parameter space Θ,
- Φ is a linear operator from $\mathcal{M}(\Theta)$ to $\mathbb{R}^T$ such that: $\Phi\mu = (\int \phi(\theta, \sigma_i)\, \mu(\mathrm{d}\theta))$, $1 \leq i \leq T$).

Recall that we first consider the estimation of one spectrum, that is $n = 1$. Thus, we have $Y \in \mathbb{R}^T$. We remark that we have for $\mu = \sum_{i=1}^K B_i \delta_{\theta_i}$ and $\vartheta = (\theta_1, \cdots, \theta_K)$:

$$\mathcal{G}_{\lambda,\varphi}(\mu) = \mathcal{F}_{\lambda,\varphi}(B, \vartheta).$$

The authors of [Boyer et al., 2019] showed that (5.5) admits an atomic solution composed of at most T atoms (or Dirac measures). Hence, when $h > 1$, $K \geq T$ and $n = 1$ in (5.3), we can build from a solution $(\hat{B}, \hat{\vartheta})$ of (5.3) a solution $\hat{\mu} = \sum_{i=1}^K \hat{B}_i \delta_{\hat{\theta}_i}$ to (5.5). Reciprocally, a solution $\tilde{\mu} = \sum_{i=1}^T \tilde{B}_i \delta_{\tilde{\theta}_i}$ of (5.5) composed of at most T atoms (there exists at least one according to [Boyer et al., 2019]) gives a solution for the problem (5.3). Therefore, provided that $K \geq T$ we may be interested in finding an atomic solution to (5.5) to solve (5.3). A major advantage of studying (5.5) is that it is a convex problem. However, the minimization is performed over an infinite dimensional space.

In practice, most of the numerical methods developed to solve (5.5) seek a solution that is atomic. We shall present in this section methods to solve (5.5) that are modifications of the celebrated Frank-Wolfe algorithm introduced in [Frank and Wolfe, 1956] (also called the Conditional Gradient Method, see [Levitin and Poljak, 1966]). This algorithm is particularly suitable for a minimization over a Banach space since it does not use at all the Hilbert structure. The Frank-Wolfe algorithm solves the optimization problem:

$$\min_{x \in \mathcal{D}} f,$$

for a (weakly) compact convex subset $\mathcal{D}$ of a Banach space E and a real valued convex differentiable function f.

It can be implemented as in Algorithm 3. It consists in using a first order approximation of the objective function around the current approximation of the minimum, minimizing the approximating function and then iterating. We remark that the second step of Algorithm 3

Algorithm 3: The Frank-Wolfe algorithm.

Input: objective function f, domain $\mathcal{D}$ and $x^{(0)} \in \mathcal{D}$ a starting point
Output: $x^{(K)}$
Initialize: $i := 0$
while $i < K$ **do**
  $y^{(i)} \in \underset{y \in \mathcal{D}}{\operatorname{argmin}} \quad f(x^{(i)}) + df(x^{(i)})[y - x^{(i)}]$ `// Step 1`
  $\alpha = \frac{2}{i+2}$ `// Step 2`
  $x^{(i+1)} = x^{(i)} + \alpha(y^{(i)} - x^{(i)})$ `// Step 3`
  $i = i + 1$
end

can be replaced by a line search. Under some smoothness asumptions on f and assuming that $\mathcal{D}$ is a subset of some reflexive Banach set E, the Frank-Wolfe algorithm satisfies weak convergence results and reaches the convergence rate (see [Levitin and Poljak, 1966, Theorem 6.1]):

$$f(x^{(k)}) - f(x^\star) = \mathcal{O}\left(1/k\right).$$

The Frank-Wolfe algorithm can be adapted to the BLasso problem which corresponds to a minimization over the non reflexive set of signed Radon measures $\mathcal{M}(\Theta)$ regularized via a total variation norm. When applied to the BLasso, it still reaches the convergence rate $\mathcal{O}(1/k)$, see [Denoyelle, 2018, Lemma 13] in this direction. A translation of the Frank-Wolfe algorithm yields iterates that are atomic, see [Boyd et al., 2017]. However, these iterates might have a large support. Recent work has taken advantage of a particularity of the Frank-Wolfe algorithm to improve the sparsity of the iterates. Indeed a fourth step can be added to the Frank-Wolfe algorithm without impacting its convergence. This step consists in replacing at the end of each iteration the approximation $x^{(k)}$ of a minimum $x^\star$ by any point x belonging to the domain $\mathcal{D}$ that does not increase the objective function, *i.e* such that $f(x) \leq f(x^{(k)})$. A line of research has recently proposed to include in the fourth step non-convex optimization problems to improve the current approximation of the minimum. Variants of the Frank-Wolfe algorithm have been proposed for minimization problems over the space of Radon measures in [Bredies and Pikkarainen, 2013], [Boyd et al., 2017] and [Denoyelle et al., 2020]. The algorithms from [Boyd et al., 2017] and [Denoyelle et al., 2020] amount to adding a Dirac measure at each iteration to approximate the data. In [Boyd et al., 2017] the optimization on the weights and positions of the Dirac measures are performed alternatively while in [Denoyelle et al., 2020] the weights and positions are modified simultaneously. We mention the interesting convergence result of [Denoyelle et al., 2020] (see Theorem 3 therein) establishing that the proposed variant of the Frank-Wolfe algorithm finds a solution of the BLasso in a finite number of steps for some admissible dictionaries provided that the BLasso admits a unique atomic solution and a non-degeneracy condition is verified. This modification is called the Sliding Frank-Wolfe (SFW) algorithm. We translate the method in Algorithm 4 to solve (5.3) when $n = 1$ and $h > 1$. We stress that a stopping criterion is used in the procedure. The algorithm can easily be adapted to the case where the linear coefficients are required to be positive, as mentionned in [Denoyelle et al., 2020]. In this case, one must remove the absolute value in the quantity involved in the stopping criterion and add positivity constraints at each minimization performed in the algorithm, see Remark 8 therein.

Let us mention that another promising off-the-grid method, called the Conic Particle Gradient Descent (CPGD), has been developed during the elaboration of this thesis. In [Chizat, 2021], convergence rates are proven for the CPGD.

We shall now get back to Problem (5.3) in the more general case where $n \geq 1$ and $h \geq 0$. The work of [Golbabaee and Poon, 2022] extended the SFW algorithm to fit several signals simultaneously. We propose a variant, see Algorithm 5 below, that separates the optimization of linear and nonlinear parameters and which merges peaks that are highly overlapping. This allows the use of classical algorithms to solve a standard group-Lasso problem for linear parameters. We note that for obvious numerical reasons, we remove before each nonlinear minimization the peaks whose linear coefficients are all zero.

Algorithm 5 can also be adapted to the problem (5.2) with positivity constraints by requiring the positivity of the linear coefficients in each linear subproblem involved in the procedure.

We developed a python library to solve (5.3) and (5.2) with Algorithm 5.

Algorithm 4: The Sliding Frank-Wolfe algorithm from [Denoyelle et al., 2020] ($n = 1$)

Data: $Y \in \mathbb{R}^T$
Input: the dictionary φ and the tuning parameter λ, K
Output: ϑ, B
Initialize: $i := 0$, $R^{(0)} := Y$ (current residuals) , $\vartheta^{(0)} := \emptyset$ (current nonlinear parameters), $N^{(0)} := 0$ (current number of peaks)
while $i < K$ **do**
  $\theta^{(i+\frac{1}{2})} \in \underset{\theta \in \Theta}{\operatorname{argmax}} \left| \left\langle R^{(i)}, \phi(\theta) \right\rangle \right|$
  $\vartheta^{(i+\frac{1}{2})} = \left(\vartheta^{(i)}, \theta^{(i+\frac{1}{2})} \right)$ and $N^{(i+1)} = N^{(i)} + 1$ `// Adding new peak`
  if $\sup_{\theta \in \Theta} \frac{2}{T\lambda} \left| \left\langle R^{(i)}, \phi(\theta) \right\rangle \right| \leq 1$ **then**
    Stop
  else
    $B^{(i+\frac{1}{2})} \in \underset{B \in \mathbb{R}^{N^{(i+1)}}}{\operatorname{argmin}} \mathcal{F}_{\lambda,\varphi}(B, \vartheta^{(i+\frac{1}{2})})$ `// Update of the linear coefficients`
    $(B^{(i+1)}, \vartheta^{(i+1)}) \in \underset{B \in \mathbb{R}^{N^{(i+1)}}, \vartheta \in \Theta^{i+1}}{\operatorname{argmin}} \mathcal{F}_{\lambda,\varphi}(B, \vartheta)$ initialized in $(B^{(i+\frac{1}{2})}, \vartheta^{(i+\frac{1}{2})})$
      `// Non-convex step`
    Remove from $\vartheta^{(i+1)}$ the coordinates $\theta_k^{(i+1)}$ such that $B_k^{(i+1)} = 0$ and update $N^{(i+1)}$
      `// Remove useless peaks`
  end
  $R^{(i+1)} = Y - B^{(i+1)}\Phi(\vartheta^{(i+1)})$
  $i = i + 1$
end

Algorithm 5: Sliding Frank-Wolfe iterations for a set of spectra ($n \geq 1$) with a stopping criteria

Data: $Y \in \mathbb{R}^{n \times T}$
Input: φ, λ, h, K
Output: ϑ, B
Initialize: $i := 0$, $R^{(0)} := Y$ (current residuals) , $\vartheta^{(0)} := \emptyset$ (current nonlinear parameters), $N^{(0)} := 0$ (current number of peaks)
while $i < K$ **do**
  $\theta^{(i+\frac{1}{2})} \in \underset{\theta \in \Theta}{\operatorname{argmax}} \left\| R^{(i)} \phi(\theta)^\top \right\|_{\ell_2}^2$
  $\vartheta^{(i+\frac{1}{2})} = \left(\vartheta^{(i)}, \theta^{(i+\frac{1}{2})} \right)$ and $N^{(i+1)} = N^{(i)} + 1$ `// Adding new peak`
  if $\sup_{\theta \in \Theta} \frac{2}{T n \lambda} \left\| R^{(i)} \phi(\theta)^\top \right\|_{\ell_2}^2 \leq 1$ **then**
    Stop
  else
    $B^{(i+\frac{1}{2})} \in \underset{B \in \mathbb{R}^{n \times (N^{(i+1)})}}{\operatorname{argmin}} \mathcal{F}_{\lambda,\varphi}(B, \vartheta^{(i+\frac{1}{2})})$ `// Update of the linear coefficients`
    Remove from $\vartheta^{(i+\frac{1}{2})}$ the coordinates $\theta_k^{(i+\frac{1}{2})}$ such that $\left\| B_{\cdot,k}^{(i+\frac{1}{2})} \right\|_{\ell_2} = 0$ and update $N^{(i+1)}$ `// Remove useless peaks`
    $\vartheta^{(i+1)} \in \underset{\vartheta \in \Theta^{i+1}}{\operatorname{argmin}} \mathcal{F}_{\lambda,\varphi}(B^{(i+\frac{1}{2})}, \vartheta)$ initialized in $\vartheta^{(i+\frac{1}{2})}$ `// Non-convex step`
    Merging routine $(\vartheta^{(i+1)}, h)$ `// Merging overlapping peaks`
    $B^{(i+1)} \in \underset{B \in \mathbb{R}^{n \times (i+1)}}{\operatorname{argmin}} \mathcal{F}_{\lambda,\varphi}(B, \vartheta^{(i+1)})$ `// Re-estimation of linear parameters`
  end
  $R^{(i+1)} = Y - B^{(i+1)} \Phi(\vartheta^{(i+1)})$
  $i = i + 1$
end

References

[Abramowitz and Stegun, 1992] Abramowitz, M. and Stegun, I. A., editors (1992). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover Publications, Inc., New York. Reprint of the 1972 edition. (p. 67, 71, 124)

[Absil et al., 2008] Absil, P.-A., Mahony, R., and Sepulchre, R. (2008). *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, NJ. With a foreword by Paul Van Dooren. (p. 38)

[Adler and Taylor, 2007] Adler, R. J. and Taylor, J. E. (2007). *Random fields and geometry*. Springer Monographs in Mathematics. Springer, New York. (p. 14)

[Aliprantis and Border, 2006] Aliprantis, C. D. and Border, K. C. (2006). *Infinite dimensional analysis*. Springer, Berlin, third edition. A hitchhiker's guide. (p. 35, 106)

[Alsmeyer and Marquardt, 2004] Alsmeyer, F. and Marquardt, W. (2004). Automatic generation of peak-shaped models. *Applied spectroscopy*, 58(8):986–994. (p. 142)

[Antonov and Nedeltcheva, 2000] Antonov, L. and Nedeltcheva, D. (2000). Resolution of overlapping uv–vis absorption bands and quantitative analysis. *Chemical Society Reviews*, 29(3):217–227. (p. 142)

[Aragoni et al., 1995] Aragoni, M. C., Arca, M., Crisponi, G., and Nurchi, V. M. (1995). Simultaneous decomposition of several spectra into the constituent Gaussian peaks. *Analytica chimica acta*, 316(2):195–204. (p. 1, 22, 142)

[Arendt et al., 2011] Arendt, W., Batty, C. J. K., Hieber, M., and Neubrander, F. (2011). *Vector-valued Laplace transforms and Cauchy problems*, volume 96 of *Monographs in Mathematics*. Birkhäuser/Springer Basel AG, Basel, second edition. (p. 35)

[Arias-Castro et al., 2011] Arias-Castro, E., Candès, E. J., and Plan, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.*, 39(5):2533–2556. (p. 114)

[Azaïs et al., 2015] Azaïs, J.-M., de Castro, Y., and Gamboa, F. (2015). Spike detection from inaccurate samplings. *Appl. Comput. Harmon. Anal.*, 38(2):177–195. (p. 29)

[Azaïs and Wschebor, 2009] Azaïs, J.-M. and Wschebor, M. (2009). *Level sets and extrema of random processes and fields*. John Wiley & Sons, Inc., Hoboken, NJ. (p. 14, 66, 82)

[Bach, 2008] Bach, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225. (p. 76)

[Baraud, 2002] Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606. (p. 114)

[Barber et al., 2017] Barber, R. F., Reimherr, M., and Schill, T. (2017). The function-on-scalar LASSO with applications to longitudinal GWAS. *Electron. J. Stat.*, 11(1):1351–1389. (p. 74)

[Beck and Teboulle, 2009] Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202. (p. 9, 75, 154)

[Bellec and Tsybakov, 2017] Bellec, P. and Tsybakov, A. (2017). Bounds on the prediction error of penalized least squares estimators with convex penalty. In *Modern problems of stochastic analysis and statistics*, volume 208 of *Springer Proc. Math. Stat.*, pages 315–333. Springer, Cham. (p. 9)

[Bernstein and Fernandez-Granda, 2019] Bernstein, B. and Fernandez-Granda, C. (2019). Deconvolution of point sources: a sampling theorem and robustness guarantees. *Comm. Pure Appl. Math.*, 72(6):1152–1230. (p. 29)

[Bernstein et al., 2020] Bernstein, B., Liu, S., Papadaniil, C., and Fernandez-Granda, C. (2020). Sparse recovery beyond compressed sensing: separable nonlinear inverse problems. *IEEE Trans. Inform. Theory*, 66(9):5904–5926. (p. 29)

[Bhaskar et al., 2013] Bhaskar, B. N., Tang, G., and Recht, B. (2013). Atomic norm denoising with applications to line spectral estimation. *IEEE Trans. Signal Process.*, 61(23):5987–5999. (p. 76)

[Bickel et al., 2009] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732. (p. 9, 28, 32, 41, 75, 83)

[Boyd et al., 2017] Boyd, N., Schiebinger, G., and Recht, B. (2017). The alternating descent conditional gradient method for sparse inverse problems. *SIAM J. Optim.*, 27(2):616–639. (p. 15, 22, 29, 76, 145, 156)

[Boyer et al., 2019] Boyer, C., Chambolle, A., De Castro, Y., Duval, V., de Gournay, F., and Weiss, P. (2019). On representer theorems and convex regularization. *SIAM J. Optim.*, 29(2):1260–1281. (p. 12, 26, 29, 76, 155)

[Boyer et al., 2017] Boyer, C., De Castro, Y., and Salmon, J. (2017). Adapting to unknown noise level in sparse deconvolution. *Inf. Inference*, 6(3):310–348. (p. 14, 17, 29, 30, 41, 47, 76, 86, 114, 144)

[Bredies and Pikkarainen, 2013] Bredies, K. and Pikkarainen, H. K. (2013). Inverse problems in spaces of measures. *ESAIM Control Optim. Calc. Var.*, 19(1):190–218. (p. 12, 15, 156)

[Brezis, 2011] Brezis, H. (2011). *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York. (p. 105)

[Bühlmann and van de Geer, 2011] Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications. (p. 2, 10, 26, 75)

[Bunea et al., 2007] Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194. (p. 9, 75)

[Butucea et al., 2021] Butucea, C., Delmas, J.-F., Dutfoy, A., and Hardy, C. (2021). Modeling infra-red spectra: an algorithm for an automatic and simultaneous analysis. In *In Proceedings of the 31st European Safety and Reliability Conference*, pages 3359–3366. (p. 1, 4, 16, 22, 29, 73, 75, 111, 113, 141)

[Butucea et al., 2022a] Butucea, C., Delmas, J.-F., Dutfoy, A., and Hardy, C. (2022a). Off-the-grid learning of sparse mixtures from a continuous dictionary. *arXiv preprint arXiv:2207.00171*. (p. 16, 22, 24, 74, 76, 77, 78, 79, 80, 81, 82, 84, 85, 86, 87, 88, 89, 90, 97, 99, 100, 101, 102, 103, 104, 106, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 126, 127, 131, 135, 136, 139, 140)

[Butucea et al., 2022b] Butucea, C., Delmas, J.-F., Dutfoy, A., and Hardy, C. (2022b). Off-the-grid prediction and testing for mixtures of translated features. *arXiv preprint arXiv:2212.01169*. (p. 19, 110)

[Butucea et al., 2022c] Butucea, C., Delmas, J.-F., Dutfoy, A., and Hardy, C. (2022c). Simultaneous off-the-grid learning of mixtures issued from a continuous dictionary. *arXiv preprint arXiv:2210.16311*. (p. 18, 72)

[Byrd et al., 1995] Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. Y. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, 16(5):1190–1208. (p. 146)

[Candès and Tao, 2007] Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351. (p. 9, 28, 75)

[Candès, 2006] Candès, E. J. (2006). Modern statistical estimation via oracle inequalities. *Acta Numer.*, 15:257–325. (p. 9)

[Candès and Davenport, 2013] Candès, E. J. and Davenport, M. A. (2013). How well can we estimate a sparse vector? *Appl. Comput. Harmon. Anal.*, 34(2):317–323. (p. 10, 29)

[Candès and Fernandez-Granda, 2013] Candès, E. J. and Fernandez-Granda, C. (2013). Super-resolution from noisy data. *J. Fourier Anal. Appl.*, 19(6):1229–1254. (p. 5, 12, 13, 14, 18, 29, 30, 41, 43, 76, 86, 87, 114, 116)

[Candès and Fernandez-Granda, 2014] Candès, E. J. and Fernandez-Granda, C. (2014). Towards a mathematical theory of super-resolution. *Comm. Pure Appl. Math.*, 67(6):906–956. (p. 4, 12, 13, 15, 18, 29, 30, 41, 75, 76, 85, 114)

[Candès and Plan, 2011] Candès, E. J. and Plan, Y. (2011). A probabilistic and RIPless theory of compressed sensing. *IEEE Trans. Inform. Theory*, 57(11):7235–7254. (p. 41, 85)

[Candes and Tao, 2005] Candes, E. J. and Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215. (p. 41)

[Candès and Wakin, 2008] Candès, E. J. and Wakin, M. B. (2008). An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30. (p. 2)

[Chen and Donoho, 1994] Chen, S. and Donoho, D. (1994). Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44 vol.1. (p. 9)

[Chen et al., 1998] Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61. (p. 9)

[Chesneau and Hebiri, 2008] Chesneau, C. and Hebiri, M. (2008). Some theoretical results on the grouped variables Lasso. *Math. Methods Statist.*, 17(4):317–326. (p. 76)

[Chizat, 2021] Chizat, L. (2021). Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, pages 1–46. (p. 15, 29, 76, 156)

[Conlon et al., 2003] Conlon, E. M., Liu, X. S., Lieb, J. D., and Liu, J. S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Sciences*, 100(6):3339–3344. (p. 1)

[De Castro et al., 2021] De Castro, Y., Gadat, S., Marteau, C., and Maugis-Rabusseau, C. (2021). SuperMix: sparse regularization for mixtures. *Ann. Statist.*, 49(3):1779–1809. (p. 12, 29, 41)

[de Castro and Gamboa, 2012] de Castro, Y. and Gamboa, F. (2012). Exact reconstruction using Beurling minimal extrapolation. *J. Math. Anal. Appl.*, 395(1):336–354. (p. 11, 13, 29, 33, 76, 114, 121, 154)

[Denoyelle, 2018] Denoyelle, Q. (2018). *Theoretical and Numerical Analysis of Super-Resolution Without Grid*. Theses, Université Paris sciences et lettres. (p. 156)

[Denoyelle et al., 2020] Denoyelle, Q., Duval, V., Peyré, G., and Soubies, E. (2020). The sliding Frank-Wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 36(1):014001, 42. (p. 15, 22, 29, 37, 75, 76, 145, 146, 156, 157)

[Diestel and Uhl, 1977] Diestel, J. and Uhl, Jr., J. J. (1977). *Vector measures*. Mathematical Surveys, No. 15. American Mathematical Society, Providence, R.I. With a foreword by B. J. Pettis. (p. 73, 105, 106)

[Donoho, 2006] Donoho, D. L. (2006). For most large underdetermined systems of equations, the minimal l_1-norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.*, 59(7):907–934. (p. 9)

[Donoho et al., 2006] Donoho, D. L., Elad, M., and Temlyakov, V. N. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1):6–18. (p. 9, 28)

[Duarte-Carvajalino and Sapiro, 2009] Duarte-Carvajalino, J. M. and Sapiro, G. (2009). Learning to sense sparse signals: simultaneous sensing matrix and sparsifying dictionary optimization. *IEEE Trans. Image Process.*, 18(7):1395–1408. (p. 3)

[Duval, 2021] Duval, V. (2021). An epigraphical approach to the representer theorem. *J. Convex Anal.*, 28(3):819–836. (p. 76)

[Duval and Peyré, 2015] Duval, V. and Peyré, G. (2015). Exact support recovery for sparse spikes deconvolution. *Found. Comput. Math.*, 15(5):1315–1355. (p. 3, 4, 12, 13, 14, 17, 28, 29, 30, 41, 75, 76, 77, 85, 111, 114, 116, 131, 138)

[Duval and Peyré, 2017a] Duval, V. and Peyré, G. (2017a). Sparse regularization on thin grids I: the Lasso. *Inverse Problems*, 33(5):055008, 29. (p. 5, 10, 28, 75, 154)

[Duval and Peyré, 2017b] Duval, V. and Peyré, G. (2017b). Sparse spikes super-resolution on thin grids II: the continuous basis pursuit. *Inverse Problems*, 33(9):095008, 42. (p. 10, 11)

[Ekanadham et al., 2011] Ekanadham, C., Tranchina, D., and Simoncelli, E. P. (2011). Recovery of sparse translation-invariant signals with continuous basis pursuit. *IEEE Trans. Signal Process.*, 59(10):4735–4744. (p. 4, 11)

[Elad and Aharon, 2006] Elad, M. and Aharon, M. (2006). Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.*, 15(12):3736–3745. (p. 3)

[Elvira et al., 2021] Elvira, C., Gribonval, R., Soussen, C., and Herzet, C. (2021). When does OMP achieve exact recovery with continuous dictionaries? *Appl. Comput. Harmon. Anal.*, 51:374–413. (p. 29)

[Ermakov, 1990] Ermakov, M. S. (1990). Minimax detection of a signal in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(4):704–715. (p. 114)

[Evans and Gariepy, 2015] Evans, L. C. and Gariepy, R. F. (2015). *Measure theory and fine properties of functions*. Textbooks in Mathematics. CRC Press, Boca Raton, FL, revised edition. (p. 67)

[Foucart and Rauhut, 2013] Foucart, S. and Rauhut, H. (2013). *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York. (p. 2, 44)

[Frank and Wolfe, 1956] Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Res. Logist. Quart.*, 3:95–110. (p. 155)

[Giné and Nickl, 2016] Giné, E. and Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York. (p. 26, 114)

[Golbabaee and Poon, 2022] Golbabaee, M. and Poon, C. (2022). An off-the-grid approach to multi-compartment magnetic resonance fingerprinting. *Inverse Problems*, 38(8):Paper No. 085002, 31. (p. 13, 14, 15, 22, 29, 77, 81, 85, 99, 142, 144, 145, 156)

[Golub and Pereyra, 1973] Golub, G. H. and Pereyra, V. (1973). The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM J. Numer. Anal.*, 10:413–432. (p. 11, 28)

[Gribonval and Nielsen, 2003] Gribonval, R. and Nielsen, M. (2003). Sparse representations in unions of bases. *IEEE Trans. Inform. Theory*, 49(12):3320–3325. (p. 3)

[Harris et al., 1994] Harris, T., Grober, R., Trautman, J., and Betzig, E. (1994). Super-resolution imaging spectroscopy. *Applied spectroscopy*, 48(1):14A–21A. (p. 5)

[Hollas, 2004] Hollas, J. M. (2004). *Modern spectroscopy*. John Wiley & Sons. (p. 22, 142, 143)

[Huang and Zhang, 2010] Huang, J. and Zhang, T. (2010). The benefit of group sparsity. *Ann. Statist.*, 38(4):1978–2004. (p. 10, 76)

[Ingster and Suslina, 2003] Ingster, Y. I. and Suslina, I. A. (2003). *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169 of *Lecture Notes in Statistics*. Springer-Verlag, New York. (p. 15, 114)

[Ingster et al., 2010] Ingster, Y. I., Tsybakov, A. B., and Verzelen, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.*, 4:1476–1526. (p. 22, 114, 128, 129)

[Jansson, 1984] Jansson, P. A. (1984). *Deconvolution: With Applications in Spectroscopy*. With Applications in Spectroscopy. Academic Press. (p. 143)

[Kaufman, 1975] Kaufman, L. (1975). A variable projection method for solving separable nonlinear least squares problems. *Nordisk Tidskr. Informationsbehandling (BIT)*, 15(1):49–57. (p. 11, 28)

[Kneip and Gasser, 1988] Kneip, A. and Gasser, T. (1988). Convergence and consistency results for self-modeling nonlinear regression. *Ann. Statist.*, 16(1):82–112. (p. 28)

[Koltchinskii, 2009] Koltchinskii, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828. (p. 9)

[Kriesten et al., 2008] Kriesten, E., Alsmeyer, F., Bardow, A., and Marquardt, W. (2008). Fully automated indirect hard modeling of mixture spectra. *Chemometrics and Intelligent Laboratory Systems*, 91(2):181–193. (p. 142)

[Lang, 1993] Lang, S. (1993). *Real and functional analysis*, volume 142 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, third edition. (p. 35)

[Laurent et al., 2012] Laurent, B., Loubes, J.-M., and Marteau, C. (2012). Non asymptotic minimax rates of testing in signal detection with heterogeneous variances. *Electron. J. Stat.*, 6:91–122. (p. 114)

[Le Gac et al., 2012] Le Gac, P.-Y., Le Saux, V., Paris, M., and Marco, Y. (2012). Ageing mechanism and mechanical degradation behaviour of polychloroprene rubber in a marine environment: Comparison of accelerated ageing and long term exposure. *Polymer degradation and stability*, 97(3):288–296. (p. 148, 149)

[Lee et al., 2006] Lee, H., Battle, A., Raina, R., and Ng, A. (2006). Efficient sparse coding algorithms. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press. (p. 3)

[Lee, 2018] Lee, J. M. (2018). *Introduction to Riemannian manifolds*, volume 176 of *Graduate Texts in Mathematics*. Springer, Cham. Second edition of [MR1468735]. (p. 37)

[Levitin and Poljak, 1966] Levitin, E. S. and Poljak, B. T. (1966). Minimization methods in the presence of constraints. *Ž. Vyčisl. Mat i Mat. Fiz.*, 6:787–823. (p. 155)

[Liu and Zhang, 2008] Liu, H. and Zhang, J. (2008). On the ℓ_1-ℓ_q regularized regression. *arXiv preprint arXiv:0802.1517*. (p. 76)

[Lounici et al., 2011] Lounici, K., Pontil, M., van de Geer, S., and Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4):2164–2204. (p. 10, 32, 76, 83, 144, 145, 147)

[Mairal et al., 2012] Mairal, J., Bach, F., and Ponce, J. (2012). Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804. (p. 3)

[Mairal et al., 2009] Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2009). Online dictionary learning for sparse coding. ICML '09, page 689–696, New York, NY, USA. Association for Computing Machinery. (p. 1)

[Mairal et al., 2008] Mairal, J., Elad, M., and Sapiro, G. (2008). Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69. (p. 1)

[Mallat, 2009] Mallat, S. (2009). *A wavelet tour of signal processing : the sparse way*. Elsevier/Academic Press, Amsterdam, third edition. With contributions from Gabriel Peyré. (p. 3, 37, 116)

[Mallat and Zhang, 1993] Mallat, S. and Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415. (p. 4, 11)

[Nardi and Rinaldo, 2008] Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Stat.*, 2:605–633. (p. 10, 76)

[Natarajan, 1995] Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234. (p. 9)

[Nocedal and Wright, 2006] Nocedal, J. and Wright, S. J. (2006). *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition. (p. 146)

[Obozinski et al., 2011] Obozinski, G., Wainwright, M. J., and Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Ann. Statist.*, 39(1):1–47. (p. 10, 106, 144, 145)

[Olshausen and Field, 1997] Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325. (p. 2, 3, 28)

[Poon et al., 2021] Poon, C., Keriven, N., and Peyré, G. (2021). The geometry of off-the-grid compressed sensing. *Foundations of Computational Mathematics*. (p. 13, 14, 17, 18, 29, 30, 37, 38, 41, 42, 43, 44, 60, 70, 76, 78, 85, 87, 109, 114, 117, 118)

[Puschmann and Kneer, 2005] Puschmann, K. G. and Kneer, F. (2005). On super-resolution in astronomical imaging. *Astronomy & Astrophysics*, 436(1):373–378. (p. 5)

[Raskutti et al., 2011] Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory*, 57(10):6976–6994. (p. 10, 29, 32, 75)

[Rice, 1944] Rice, S. O. (1944). Mathematical analysis of random noise. *Bell System Tech. J.*, 23:282–332. (p. 14)

[Rubinstein et al., 2010] Rubinstein, R., Bruckstein, A. M., and Elad, M. (2010). Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057. (p. 3)

[Sakai, 1996] Sakai, T. (1996). *Riemannian geometry*, volume 149 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI. Translated from the 1992 Japanese original by the author. (p. 37)

[Santosa and Symes, 1986] Santosa, F. and Symes, W. W. (1986). Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Statist. Comput.*, 7(4):1307–1330. (p. 9)

[Schiebinger et al., 2018] Schiebinger, G., Robeva, E., and Recht, B. (2018). Superresolution without separation. *Inf. Inference*, 7(1):1–30. (p. 29, 43)

[Tang, 2015] Tang, G. (2015). Resolution limits for atomic decompositions via markov-bernstein type inequalities. In *2015 International Conference on Sampling Theory and Applications (SampTA)*, pages 548–552. (p. 43)

[Tang et al., 2013a] Tang, G., Bhaskar, B. N., and Recht, B. (2013a). Sparse recovery over continuous dictionaries-just discretize. In *2013 Asilomar Conference on Signals, Systems and Computers*, pages 1043–1047. IEEE. (p. 8, 28, 145, 153)

[Tang et al., 2015] Tang, G., Bhaskar, B. N., and Recht, B. (2015). Near minimax line spectral estimation. *IEEE Trans. Inform. Theory*, 61(1):499–512. (p. 12, 14, 17, 29, 30, 32, 41, 47, 76, 86, 114)

[Tang et al., 2013b] Tang, G., Bhaskar, B. N., Shah, P., and Recht, B. (2013b). Compressed sensing off the grid. *IEEE Trans. Inform. Theory*, 59(11):7465–7490. (p. 4, 12)

[Tchalla et al., 2017] Tchalla, S. T., Le Gac, P.-Y., Maurin, R., and Creac'Hcadec, R. (2017). Polychloroprene behaviour in a marine environment: Role of silica fillers. *Polymer Degradation and Stability*, 139:28–37. (p. 23, 149)

[Tchalla, 2017] Tchalla, T. S. (2017). *Durabilité d'assemblages métal/élastomère en milieu marin*. PhD thesis, Université de Bretagne occidentale - Brest, France. (p. 5, 149)

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288. (p. 9, 26, 28, 75)

[Tropp, 2004] Tropp, J. A. (2004). Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242. (p. 44)

[Tsybakov, 2009] Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. (p. 26)

[van de Geer, 2016] van de Geer, S. (2016). *Estimation and testing under sparsity*, volume 2159 of *Lecture Notes in Mathematics*. Springer, [Cham]. Lecture notes from the 45th Probability Summer School held in Saint-Four, 2015, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. (p. 9, 28, 41)

[van de Geer and Bühlmann, 2009] van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392. (p. 41, 75)

[Wainwright, 2019] Wainwright, M. J. (2019). *High-dimensional statistics*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. A non-asymptotic viewpoint. (p. 2)

[Yaghoobi et al., 2009] Yaghoobi, M., Daudet, L., and Davies, M. E. (2009). Parametric dictionary design for sparse coding. *IEEE Trans. Signal Process.*, 57(12):4800–4810. (p. 3)

[Yuan and Lin, 2006] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67. (p. 10, 76, 144)

[Zhu et al., 1997] Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Software*, 23(4):550–560. (p. 146)