



# Propriétés fréquentistes des méthodes Bayésiennes semi-paramétriques et non paramétriques

Jean-Bernard Salomond

## ► To cite this version:

Jean-Bernard Salomond. Propriétés fréquentistes des méthodes Bayésiennes semi-paramétriques et non paramétriques. Mathématiques générales [math.GM]. Université Paris Dauphine - Paris IX, 2014. Français. NNT : 2014PA090034 . tel-01087106

**HAL Id: tel-01087106**

**<https://theses.hal.science/tel-01087106>**

Submitted on 25 Nov 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Année : 2014

THÈSE  
présentée à  
L'UNIVERSITÉ PARIS-DAUPHINE  
ÉCOLE DOCTORALE DE DAUPHINE  
Centre de **R**echerche en **M**athématiques de la **D**écision  
Pour l'obtention du titre de  
DOCTEUR EN MATHÉMATIQUES APPLIQUÉES  
Présentée et soutenue par  
Jean-Bernard SALOMOND  
Le 30 Septembre 2014

---

Propriétés fréquentistes des méthodes  
Bayésiennes semi-paramétriques et  
non-paramétriques

---

**Jury :**

Ismaël CASTILLO	CNRS	<i>Examineur</i>
Fabienne COMTE	Université Paris Descartes	<i>Examineur</i>
Cécile DUROT	Université Paris Ouest Nanterre La Défense	<i>Examineur</i>
Elisabeth GASSIAT	Université Paris-Sud	<i>Rapporteur</i>
Dominique PICARD	Université Paris Diderot - Paris 7	<i>Examineur</i>
Vincent RIVOIRARD	Université Paris Dauphine	<i>Examineur</i>
Judith ROUSSEAU	Université Paris-Dauphine	<i>Directrice de thèse</i>
Harry VAN ZANTEN	University of Amsterdam	<i>Rapporteur</i>



# Résumé

La recherche sur les méthodes bayésiennes non-paramétriques connaît un essor considérable depuis les vingt dernières années notamment depuis le développement d'algorithmes de simulation permettant leur mise en pratique. Il est donc nécessaire de comprendre, d'un point de vue théorique, le comportement de ces méthodes. Cette thèse présente différentes contributions à l'analyse des propriétés fréquentistes des méthodes bayésiennes non-paramétriques. Si se placer dans un cadre asymptotique peut paraître restrictif de prime abord, cela permet néanmoins d'appréhender le fonctionnement des procédures bayésiennes dans des modèles extrêmement complexes. Cela permet notamment de détecter les aspects de l'a priori particulièrement influents sur l'inference. De nombreux résultats généraux ont été obtenus dans ce cadre, cependant au fur et à mesure que les modèles deviennent de plus en plus complexes, de plus en plus réalistes, ces derniers s'écartent des hypothèses classiques et ne sont plus couverts par la théorie existante. Outre l'intérêt intrinsèque de l'étude d'un modèle spécifique ne satisfaisant pas les hypothèses classiques, cela permet aussi de mieux comprendre les mécanismes qui gouvernent le fonctionnement des méthodes bayésiennes non-paramétriques.

**Chapitre 1** L'introduction présente le paradigme bayésien et l'approche bayésienne des problèmes non-paramétriques. Nous introduisons les propriétés fréquentistes des méthodes bayésiennes et présentons leur importance dans la compréhension du comportement de la loi a posteriori. Nous présentons ensuite les principaux modèles étudiés dans cette thèse, et les difficultés posées par ceux-ci pour l'étude de leurs propriétés asymptotiques.

**Chapitre 2** Dans ce chapitre, nous étudions la consistance et la vitesse de concentration de la loi a posteriori dans le modèle de densité décroissante pour différentes métriques. Ce modèle est particulièrement intéressant car les densités décroissantes ont une représentation sous forme de mélange d'uniformes et sont donc un cas particulier de mélange pour lequel le support du noyau dépend du paramètre. Dans ce cadre, les hypothèses classiques nécessaires pour la consistance de la loi a posteriori ne sont pas vérifiées. Notamment la loi a priori ne met pas suffisamment de masse sur les voisinages de Kullback-

Leibler du vrai paramètre, et une adaptation des méthodes usuelles est donc nécessaire. Pour deux familles d'a priori classiques, nous prouvons que l'a posteriori se concentre à la vitesse maximax pour les pertes  $L_1$  et Hellinger. Nous étudions ensuite la consistance de la loi a posteriori de la densité pour les pertes ponctuelle et norme sup. Ces deux métriques sont en général difficiles à étudier car elles ne peuvent être reliées à la divergence naturelle qu'est la divergence de Kullback-Leibler. Pour ces deux pertes, nous prouvons la consistance de l'a posteriori et donnons une borne supérieure pour la vitesse de concentration.

**Chapitre 3** Nous proposons un test bayésien non paramétrique de décroissance d'une fonction dans le modèle de régression gaussien. Dans ce cadre, outre le fait que les deux hypothèses sont non-paramétriques, l'hypothèse nulle est incluse dans l'alternative. Il s'agit donc d'un cas de test particulièrement difficile. En outre dans ce cas, l'approche usuelle par le facteur de Bayes n'est pas consistante. Nous proposons donc une approche alternative reprenant les idées d'approximation d'une hypothèse ponctuelle par un intervalle. Nous prouvons que pour une large famille de lois a priori, le test proposé est consistant et sépare les hypothèses à la vitesse maximax. De plus notre procédure est facile à implémenter et à mettre en œuvre. Nous étudions ensuite son comportement sur des données simulées et comparons les résultats avec les méthodes classiques existantes dans la littérature. Pour chacun des cas considérés, nous obtenons des résultats au moins aussi bons que les méthodes existantes, et les surpassons pour un certain nombre de cas.

**Chapitre 4** (co-écrit avec Bartek Knapik) Nous proposons une méthode générale pour l'étude des problèmes inverses linéaires mal-posés dans un cadre bayésien. S'il existe de nombreux résultats sur les méthodes de régularisation et la vitesse de convergence d'estimateurs classiques, pour l'estimation de fonctions dans un problème inverse mal-posé, les vitesses de concentration d'a posteriori dans le cadre bayésien n'a été que très peu étudié dans ce cadre. De plus ces quelques rares résultats existant ne considèrent que des familles très limitées de lois a priori, en général reposant sur la décomposition en valeurs singulières de l'opérateur considéré. Dans ce chapitre nous proposons des conditions générales sur la loi a priori sous lesquelles l'a posteriori se concentre à une certaine vitesse. Notre approche nous permet de trouver les vitesses de concentration de l'a posteriori pour de nombreux modèles et de larges classes de loi a priori. Cette approche est de plus particulièrement intéressante car elle permet de mieux comprendre le fonctionnement de la loi a posteriori et notamment l'impact de l'opérateur sur l'inférence.





# Summary

Research on Bayesian nonparametric methods has received a growing interest for the past twenty years, especially since the development of powerful simulation algorithms which makes the implementation of complex Bayesian methods possible. From that point it is necessary to understand from a theoretical point of view the behaviour of Bayesian nonparametric methods. This thesis presents various contributions to the study of frequentist properties of Bayesian nonparametric procedures. Although studying these methods from an asymptotic angle may seem restrictive, it allows to grasp the operation of the Bayesian machinery in extremely complex models. Furthermore, this approach is particularly useful to detect the characteristics of the prior that are strongly influential in the inference. Many general results have been proposed in the literature in this setting, however the more complex and realistic the models the further they get from the usual assumptions. Thus many models that are of great interest in practice are not covered by the general theory. If the study of a model that does not fall under the general theory has an interest on its own, it also allows for a better understanding of the behaviour of Bayesian nonparametric methods in a general setting.

**Chapter 1** The introduction presents the Bayesian paradigm and the Bayesian approach to nonparametric problems. We introduce frequentist properties of Bayesian procedures and present their importance in the understanding of the behaviour of the posterior distribution. We then present the different models studied in this manuscript and the challenge faced in studying of their asymptotic properties.

**Chapter 2** In this chapter, we study consistency and concentration rates of the posterior distribution under several metrics in the monotone density model. This model is particularly interesting as monotone densities can be written as a mixture of uniform kernels which is a special case of kernels for which the support depends on the parameter. In this case the usual hypotheses required to derive posterior concentration rate are not satisfied. In particular, the prior distribution we consider do not put positive mass on Kullback-Leibler neighbourhoods of the true parameter and we thus have to adapt the

standard methods to get an upper bound on the posterior concentration rate. For two standard prior distributions, we prove that the posterior concentrate at the minimax rate for the  $L_1$  and the Hellinger losses. We then study consistency of the posterior under the pointwise and supremum loss. These two metrics are in general difficult to study in the Bayesian framework as they are not related to the Kullback-Leibler divergence which is the natural semi-metric in this setting. We however prove that the posterior is consistent for both losses and get an upper bound for the posterior concentration rate.

**Chapter 3** We propose a Bayesian nonparametric procedure to test for monotonicity in the regression setting. In this case, not only the null and the alternative hypotheses are nonparametric, but one is embedded in the other which makes the testing problem particularly difficult. In particular the Bayes-Factor, which is a usual Bayesian answer to testing problems, is not consistent under the null hypothesis. We propose an alternative approach that relies on the ideas of approximating a point null hypothesis by shrinking intervals. The proposed procedure is consistent for a wide family of prior distributions and separate the hypotheses at the minimax rate. Furthermore, our approach is easy to implement and does not require heavy computations contrariwise to the existing procedures. We then study its behaviour on simulated data and for all the considered cases, our procedure does at least as good as the classical ones, and outperform them in some cases.

**Chapter 4** (Joint work with Bartek Knapik) We propose a general approach to study nonparametric ill-posed linear inverse problems in a Bayesian setting. Although there is a wide literature on regularisation methods and convergence of estimators in this setting, the posterior concentration in a Bayesian setting has not received much attention yet. Furthermore, the few existing results only consider very restricted families of prior distributions, mostly related to the singular value decomposition of the operator at hand. In this chapter we give general conditions on the prior such that the posterior concentrates at a certain rate. This approach allows us to derive asymptotic results for various ill-posed inverse problems and wide families of priors. Furthermore, this approach is particularly interesting in the sense that it gives some valuable insights on the behaviour of the posterior distribution in these models and the impact on the operator on the inference.

# Contents

<b>Résumé</b>	<b>iii</b>
<b>Summary</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Bayesian nonparametric approaches . . . . .	2
1.1.1 Bayesian modeling . . . . .	2
1.1.2 Bayesian nonparametrics . . . . .	3
1.2 Asymptotic properties of the posterior distribution . . . . .	5
1.2.1 Posterior consistency . . . . .	5
1.2.2 Posterior concentration rate . . . . .	7
1.2.3 Minimax concentration rates and adaptation . . . . .	9
1.3 Nonparametric Bayesian testing . . . . .	10
1.4 Challenging asymptotic properties . . . . .	11
1.4.1 Inference under monotonicity constraints . . . . .	12
1.4.2 Ill posed linear inverse problems . . . . .	15
<b>2 Monotone densities</b>	<b>27</b>
2.1 Introduction . . . . .	28
2.2 Main results . . . . .	31
2.2.1 $L_1$ and Hellinger metric . . . . .	32
2.2.2 Pointwise and supremum loss . . . . .	33
2.3 Proofs . . . . .	35
2.3.1 Proof of Theorems 2.1 and 2.2 . . . . .	35
2.3.2 Proof of Theorems 2.3 and 2.5 . . . . .	39
2.3.3 Proof of Theorem 2.4 . . . . .	42
2.3.4 Proof of Theorem 2.6 . . . . .	43
2.4 Technical Lemmas . . . . .	44
2.4.1 Proof of Lemma 2.1 . . . . .	44
2.4.2 Proof of Lemma 2.2 . . . . .	50
2.5 Additionnal proof . . . . .	51

2.6	Discussion . . . . .	53
<b>3</b>	<b>Bayesian testing for monotonicity</b>	<b>57</b>
3.1	Introduction . . . . .	58
3.1.1	Modelling with monotone constraints . . . . .	58
3.1.2	The Bayes factor approach . . . . .	59
3.1.3	An alternative approach . . . . .	59
3.2	Construction of the test . . . . .	61
3.2.1	The testing procedure . . . . .	61
3.2.2	Theoretical results . . . . .	62
3.2.3	A choice for the prior in the non informative case . . . . .	64
3.3	Simulated Examples . . . . .	65
3.4	Application to Global Warming data . . . . .	68
3.5	Proof of Theorem 3.1 . . . . .	69
3.6	Proofs of technical lemmas . . . . .	71
3.6.1	Proof of Lemma 3.1 . . . . .	71
3.6.2	Proof of lemma 3.2 . . . . .	74
3.6.3	Proof of Lemma 3.3 . . . . .	77
3.7	Discussion . . . . .	77
<b>4</b>	<b>Ill-posed inverse problems</b>	<b>81</b>
4.1	Introduction . . . . .	82
4.2	General Theorem . . . . .	83
4.3	Modulus of continuity . . . . .	85
4.4	Some models . . . . .	87
4.4.1	White noise . . . . .	87
4.4.2	Regression . . . . .	96
4.5	Discussion . . . . .	106

# Chapter 1

## Introduction

“Perhaps I should not have been a fisherman, he thought.  
But that was the thing that I was born for.”  
– **Ernest Hemingway**, *The old man and the sea*.

### Résumé

L’introduction présente le paradigme bayésien et l’approche bayésienne des problèmes non-paramétriques. Nous introduisons les propriétés fréquentistes des méthodes bayésiennes et présentons leur importance dans la compréhension du comportement de la loi a posteriori. Nous présentons ensuite les principaux modèles étudiés dans cette thèse, et les difficultés posées par ceux-ci pour l’étude de leurs propriétés asymptotiques.

This introduction presents the main concepts common to the following chapters, the statistical modeling and its Bayesian approach that we adopt in this thesis. We proceed with a quick introduction to nonparametric statistics and the construction of prior distributions in an infinite dimensional space, and we emphasize the importance of frequentists properties of Bayesian nonparametric procedures. We then present the different statistical models studied in this manuscript.

## 1.1 Bayesian nonparametric approaches

The main goal of statistics is to infer on a random phenomenon given observations. The core concept of statistics is *probabilistic modelling*, that is a mathematical approximation of the random phenomenon at hand. In a statistical model, an observation  $X$  on an observation space  $\mathcal{X}$  is assumed to be generated from a probability distribution  $P$  that belongs to a model  $\mathcal{P}$ . Usually this distribution is characterized by a parameter  $\theta$  in a parameter set  $\Theta$  which gives the *sampling model*

$$\{\mathcal{X}, P_\theta, \theta \in \Theta\}.$$

The aim of statistics is then to infer, and make decisions on the model, based on the observed data. To model complex data generating phenomenon, the parameter space  $\Theta$  may be very large and possibly infinite dimensional. As often in mathematical sciences, it is interesting to delineate regions of statistical methodology, and modern mathematical statistics tends to differentiate Bayesian versus frequentist methods, parametric versus nonparametric models. In this section, we define Bayesian nonparametric models and underline their importance.

### 1.1.1 Bayesian modeling

Statistic models usually fall into either the frequentist paradigm or the Bayesian one. The frequentist paradigm considers that the data are generated from a fixed distribution  $P_{\theta_0}$  associated with the *true* parameter  $\theta_0$ . Let  $g$  be a function from  $\Theta$  to  $\Xi$ , such that one is interested in making inference on  $g(\theta_0)$ . Frequentist statisticians look for statistics, that is functions  $S : \mathcal{X} \mapsto \Xi$  that minimizes a risk

$$R(g(\theta_0), S(X)).$$

The risk is most of the time associated with a metric or semi-metric  $d$ , or more generally any loss function, and can be rewritten

$$R(g(\theta), S(X)) = \mathbb{E}_{\theta_0} [d(g(\theta_0), S(X))],$$

where  $\mathbb{E}_\theta$  is the expectation with respect to  $P_\theta$ .

In the Bayesian paradigm, one models the ignorance on the parameter  $\theta$  through a probability distribution  $\Pi$  based on prior beliefs (the *prior distribution*). An extensive introduction to Bayesian statistics can be found in Robert (2007). A Bayesian model is thus a sampling model

$$X \sim P_\theta, \theta \in \Theta$$

together with a prior model

$$\theta \sim \Pi$$

which can be combined through the Bayes' rules to get a probability distribution of the parameter given the data called the *posterior distribution* defined as for all measurable  $A \subset \Theta$

$$\Pi(\theta \in A|X) = \frac{\int_A P_\theta(X) \Pi(d\theta)}{\int_\Theta P_\theta(X) \Pi(d\theta)}. \quad (1.1)$$

It is the single object on which all inference (e.g. estimation, testing, construction of credible sets, etc.) is based.

The Bayesian approach to statistics has become increasingly popular, especially since the 1990's because of the development of new sampling methods such as Markov-Chains Monte-Carlo (MCMC) algorithms that makes sampling under the posterior distribution feasible if not easy. Bayesian methods are now used in a wide variety of domains, from biology to finance and data analysts are more and more attracted by its axiomatic view of uncertainty and its capacity to handle complex models, see Gelman et al. (2004) for instance. However, the fact that some methods are called *Bayesian* emphasizes the fact that there is still two philosophical approaches to statistical modeling. When the parameter space is finite dimensional, Bayesian and frequentist methods usually agree when the amount of information grows. In particular, under weak assumptions on the prior distribution, the so called Bernstein-von-Mise Theorem, as presented in Le Cam and Yang (2000), shows that Bayesian credible sets and frequentist confidence intervals are asymptotically equivalent. This result is particularly important as it indicates that Bayesian models with different priors<sup>1</sup> will eventually agree when the amount of information<sup>2</sup> grows, and will give a similar answer as frequentists ones.

### 1.1.2 Bayesian nonparametrics

Nonparametric models are often defined as probabilistic models with *massively many parameters* (see Müller and Mitra, 2013) or with an infinite dimensional

---

<sup>1</sup>With a slight abuse of notations, we may say prior for prior distributions when there is no confusion

<sup>2</sup>We will call *amount of information* either the number of data points or the level of noise.

parameter space as in Ghosh and Ramamoorthi (2003). These models offer more flexibility than parametric methods but their mathematical complexity is in general more involved than for parametric methods.

A first problem in Bayesian nonparametrics is to define a probability distribution on an infinite dimensional space. Choosing a prior distribution is a key point in of the Bayesian inference, and going from prior knowledge to a prior distribution can be challenging. In particular for infinite dimensional parameter spaces, assuring that the prior distribution has a sufficiently large support is a difficult task, not mentioning the difficulty to compute the posterior distribution for such models. A popular tool in the Bayesian nonparametric literature is the Dirichlet process introduced by Ferguson (1974). The Dirichlet process is a probability measures on the set of probability measure and can be defined as follows:

**Definition 1.1** (Dirichlet process, Ferguson, 1974). Let  $\alpha$  be a non null finite measure on  $\mathcal{X}$ . We say that  $P$  follows a Dirichlet process  $DP(\alpha)$ , if for all  $k \in \mathbb{N}^*$ , all partition of measurable sets  $(B_1, \dots, B_k)$  of  $\mathcal{X}$ ,

$$(P(B_1), \dots, P(B_k)) \sim \mathcal{D}(\alpha(B_1), \dots, \alpha(B_k))$$

where  $\mathcal{D}$  is the Dirichlet distribution.

The Dirichlet processes have been proved to have a large weak support (see Ferguson, 1973), which is all distributions whose support is included in the support of the base measure  $\alpha$ . Its hyperparameters are easily interpretable and it lead to tractable posteriors. Moreover Sethuraman (1994) showed that the Dirichlet process can be obtained in a constructive way called the *stick breaking* representation. In addition, it opened the way to more flexible prior distributions on the set of density functions. Since then many prior distributions on infinite dimensional sets have been proposed. For instance Antoniak (1974) introduced mixtures of Dirichlet process in the context of probability densities estimation. Given a collection of kernels  $K_\mu(\cdot)$  depending on a parameter  $\mu$  we define the mixture

$$\theta(\cdot) = \int_{\mathcal{X}} K_\mu(\cdot) dP(\mu),$$

where  $P$  is a probability measure. Thus, choosing a prior on  $P$  (e.g. a Dirichlet process prior) induces a prior on  $\theta$ .

Many other priors have been proposed in the literature, general classes of mixtures for the density model (Lo, 1984), hierarchical Dirichlet processes (Teh et al., 2006), Gaussian processes (Lenk, 1991) among others. It is typically difficult in nonparametric settings to quantify the impact of a prior distribution on the posterior inference. If in the parametric case, the Bernstein-von-Mise theorem shows

that when the amount of information increases, inference based on different priors will merge, it does not hold easily when the parameter space is very large. Diaconis and Freedman (1986) showed that in some cases, Bayesian nonparametric procedures can lead to inconsistent results (when the data are assumed to be sampled from a distribution  $P_{\theta_0}$ , the posterior distribution does not accumulate its mass around the *true* parameter). Some other examples show that even if the posterior concentrates its mass around the true parameter, the prior still influences the rate at which this concentration occurs.

## 1.2 Asymptotic properties of the posterior distribution

Looking at the asymptotic behaviour of the posterior distribution helps understanding the impact of the prior on the posterior distribution. It is also important to detect which parts of the prior influence the most the posterior. In particular, some aspects of the prior may remain when the amount of information grows to infinity and may thus be highly influential for small sample sizes for instance. We now define two main asymptotic properties of the posterior distribution studied in this manuscript, namely posterior consistency and posterior concentration rate.

### 1.2.1 Posterior consistency

Consistency of the posterior distribution can be considered as a least requirement for Bayesian nonparametric procedures. Diaconis and Freedman (1986) proved that in the case of exchangeable data, consistency of the posterior distribution is equivalent to weak merging of posteriors associated with different proper prior distributions. This is particularly interesting as, as argued before, it is often difficult to go from prior knowledge on the parameter to a prior distribution, and two statisticians could come with two different priors. We give a more detailed definition of consistency of the posterior distribution, as presented in Ghosh and Ramamoorthi (2003).

Let the observations  $\mathbf{X}^n \in \mathcal{X}^n$  be some random variables sampled from a distribution  $P_{\theta}^n$  for  $\theta \in \Theta$ . Here  $n$  is considered to be a quantification of the amount of information. Consider  $\Pi$  a prior probability distribution on  $\Theta$ . We can thus compute the posterior distribution of  $\theta$  denoted  $\Pi(\cdot|\mathbf{X}^n)$  (see (1.1)). Assume that there exists an unknown parameter  $\theta_0 \in \Theta$  such that the data are generated from the *true* distribution  $P_{\theta_0}^n$ , and define an  $\epsilon$ -neighbourhood of  $\theta_0$  associated with the loss function  $d$

$$B_{\epsilon}(\theta_0) = \{\theta, d(\theta, \theta_0) \leq \epsilon\}.$$

**Definition 1.2.** The posterior distribution is said to be consistent at  $\theta_0$  for the loss  $d$  if for any  $\epsilon > 0$ , the posterior probability of  $B_\epsilon(\theta_0)$

$$\Pi(B_\epsilon(\theta_0)|\mathbf{X}^n) \rightarrow 1$$

either in  $P_{\theta_0}^n$  probability or  $P_{\theta_0}^\infty$ -almost surely.

A first result of Doob (1949) shows that when  $d$  is a metric and  $(\Theta, d)$  is a complete separable space, any posterior distribution is consistent at  $\theta_0$ ,  $\Pi$ -almost surely, under some ergodicity conditions. This result is interesting but very weak as it does not provides any information on the set of parameters at which consistency holds.

A usual requirement for the posterior to be consistent is that the prior puts positive mass on neighbourhoods of  $\theta_0$ . More precisely, if  $P_\theta$  is absolutely continuous with respect to  $P_{\theta_0}$ , define the Kullback-Leibler divergence as

$$KL(P_\theta, P_{\theta_0}) = \int \log \left( \frac{dP_\theta}{dP_{\theta_0}} \right) dP_\theta,$$

one will require that  $\Pi(KL(P_\theta, P_{\theta_0}) < \epsilon) > 0$  for all  $\epsilon$ .

A second condition is that the model makes it possible to differentiate between  $\theta_0$  and parameters outside  $B_\epsilon(\theta_0)$ . This can be formalized by the existence of a sequence of tests of

$$H_0 : \theta = \theta_0, \text{ versus } H_1 : \theta \in B_\epsilon^c(\theta_0). \quad (1.2)$$

We then define an exponentially consistent sequence of tests  $\{\phi_n(\mathbf{X}^n)\}$  as follows

**Definition 1.3.** The sequence of tests  $\{\phi_n(\mathbf{X}^n)\}$  is exponentially consistent for testing (1.2) if there exists  $c > 0$  such that for all  $n$

$$\mathbb{E}_{\theta_0}(\phi_n(\mathbf{X}^n)) \lesssim e^{-cn}, \quad \sup_{\theta \in B_\epsilon^c(\theta_0)} \mathbb{E}_\theta(1 - \phi_n(\mathbf{X}^n)) \lesssim e^{-cn}.$$

For independent identically distributed observations  $\mathbf{X}^n = (X_1, \dots, X_n)$  where the parameter of interest is the common density  $f$  with respect to a measure  $\lambda$ , we thus have

$$f = \frac{dP_\theta}{d\lambda}, \quad \theta^n(\mathbf{X}^n) = \prod_{i=1}^n \theta(X_i),$$

hence, in this case  $\theta = f$ , Schwartz (1965) gives general conditions on the model to achieve consistency. In this case the Kullback-Leibler divergence between  $f$  and  $f_0$  is

$$KL(f, f_0) = \int_{\mathcal{X}} f(x) \log \left( \frac{f(x)}{f_0(x)} \right) dx. \quad (1.3)$$

Schwartz (1965) requires that the prior has positive mass on all  $\epsilon$ -neighborhoods for the Kullback-Leibler divergence for all  $\epsilon > 0$

$$\Pi(f : KL(f, f_0) \leq \epsilon) > 0.$$

The truth  $f_0$  is then said to belong to the  $KL$ -support of the prior  $\Pi$ . This condition ensures that the support of the prior is large in the sense of the Kullback-Leibler divergence.

In the density setting, Schwartz's Theorem then states:

**Theorem 1.1** (Schwartz (1965)). *Let  $\Pi$  be a prior on  $\Theta$ , and  $\theta_0 \in \Theta$  such that*

- $\theta_0$  is in the  $KL$ -support of  $\Pi$
- there exists an exponentially consistent sequence of tests for (1.2)

*then  $\Pi(B_\epsilon(\theta_0)|\mathbf{X}^n) \rightarrow 1$   $P_{\theta_0}^\infty$  almost surely.*

Since this result of Schwartz, other types of results have been obtained, in many different settings, see for instance Walker and Hjort (2001), Walker (2003), Walker (2004), Lijoi et al. (2007).

### 1.2.2 Posterior concentration rate

A more refined asymptotic property is the posterior concentration rate. Loosely speaking, it is the rate at which the  $\epsilon$ -neighborhoods  $B_\epsilon(\theta_0)$  can shrink such that the posterior probability of  $B_\epsilon(\theta_0)$  remains close to 1. To get a better understanding of the impact of the prior on the posterior, we need to study sharper results than mere consistency. Some aspects of the prior may influence significantly the posterior concentration rate. They are thus likely to be highly influential for finite datasets and should thus be handled with care. We now give a precise definition of the posterior concentration rate and present some general results proposed in the literature.

**Definition 1.4.** Let the observations  $\mathbf{X}^n$  be sampled from a distribution  $P_{\theta_0}^n$  with  $\theta_0 \in \Theta$  and let  $\Pi$  be a prior on  $\Theta$ . A posterior concentration rate at  $\theta_0$  with respect to a semimetric  $d$  on  $\Theta$  is a sequence  $\epsilon_n$  such that for all positive sequences  $M_n$  going to infinity

$$\Pi(\theta, d(\theta, \theta_0) \leq M_n \epsilon_n | \mathbf{X}^n) \rightarrow 1,$$

in  $P_{\theta_0}^n$  probability as  $n$  goes to infinity.

In their seminal papers Ghosal et al. (2000a) (see also Shen and Wasserman, 2001) proposed general conditions on the model to derive posterior concentration

rates in the density model (i.e. independent and identically distributed observations  $\mathbf{X}^n$ ) in a similar way Schwartz (1965) did for consistency. This idea has then been extended to many other models, and other approaches have been proposed, see for instance Ghosal and van der Vaart (2007). Their approach requires also that the prior puts enough mass on shrinking Kullback-Leibler neighbourhoods of the truth. However the neighbourhoods here are more restrictive than the ones considered for consistency. Define the  $k$ -th centred Kullback-Leibler moment, if  $dP_\theta^n$  is absolutely continuous with respect to  $dP_{\theta_0}^n$ ,

$$V_k(P_\theta^n, P_{\theta_0}^n) = \int_{\mathcal{X}} \left| \log \left( \frac{dP_\theta^n}{dP_{\theta_0}^n} \right) - KL(P_\theta^n, P_{\theta_0}^n) \right|^k dP_\theta.$$

We then define the following Kullback-Leibler neighborhood

$$S_n(\theta_0, \epsilon, k) = \{ KL(P_\theta^n, P_{\theta_0}^n) \leq n\epsilon^2, V_k(P_\theta^n, P_{\theta_0}^n) \leq n\epsilon \}.$$

As in Schwartz's Theorem, Ghosal and van der Vaart (2007) also requires the existence of an exponentially consistent sequence of tests, but instead of testing against the complement of the shrinking ball  $B_{\epsilon_n}(\theta_0)$ , it is sufficient to test against sets

$$B_n^j(\theta_0) = \{ \theta \in \Theta_n, j\epsilon_n \leq d(\theta, \theta_0) \leq 2j\epsilon_n \},$$

for any integer  $j \geq J_0$  for some positive  $J_0$ , where  $\Theta_n$  is an increasing sequence of sets that takes most of the prior mass of  $\Pi$ . Their Theorem is thus as follows:

**Theorem 1.2** (Theorem 3 of Ghosal and van der Vaart (2007)). *Let  $d$  be a semimetric on  $\Theta$  and consider a sequence  $\epsilon_n$  such that  $\epsilon_n \rightarrow 0$ ,  $n\epsilon_n^2 \rightarrow \infty$  as  $n \rightarrow \infty$ . For  $k > 1$ ,  $K > 0$  and  $\Theta_n \subset \Theta$ , if there exists a sequence of tests  $\phi_n$  such that for  $J_0 > 0$ , for every  $j \geq J_0$*

$$E_{\theta_0} \phi_n \rightarrow 0, \quad \sup_{B_n^j(\theta_0)} (1 - \phi_n) \leq e^{-Knj^2\epsilon_n^2}, \quad (1.4)$$

and if

$$\frac{\Pi(B_n^j(\theta_0))}{\Pi(S_n(\theta_0, \epsilon_n, k))} \leq e^{Knj^2\epsilon_n^2/2}, \quad (1.5)$$

then for every sequence  $M_n \rightarrow \infty$  we have

$$\Pi(\theta \in \Theta_n, d(\theta, \theta_0) \leq M_n \epsilon_n | \mathbf{X}^n) \rightarrow 1$$

in  $P_{\theta_0}^n$ -probability as  $n$  goes to infinity.

A usual way of insuring the existence of tests in condition (1.4), for well suited semimetric  $d$  is to control the covering number of the sets  $B_n^j(\theta_0)$ . For instance when the semimetric  $d$  is the Hellinger metric, the well known results by Le Cam (1986) or Le Cam and Yang (2000) insure the existence of such sequence of tests under some entropy conditions.

### 1.2.3 Minimax concentration rates and adaptation

The concentration rate's theory can be related to the classical optimal *convergence rate* of estimators. Ghosal et al. (2000a) show in the context of density estimation, that the posterior yields a point estimate that converges at the same rate as the posterior concentration rate when the considered loss is bounded and convex. It thus makes sense to compare frequentists and Bayesian approaches based on this asymptotic property.

To study the asymptotic behaviour of the posterior distribution, we only consider some subspace  $\Theta_0$  of the parameter space on which the functions are *behaving well*. One of the most common criterion for studying optimality of an estimator is the minimax risk defined by the minimum over all estimator of maximal risk of this estimator. More precisely, if  $d$  is a semimetric on  $\Theta$ , the minimax risk over  $\Theta_0 \subset \Theta$  is defined as (see Tsybakov, 2009)

$$R_n = \inf_{T_n} \sup_{\theta \in \Theta_0} E_\theta [d(T_n, \theta)],$$

where the infimum is taken over *all estimators*  $T_n$ . The minimax rate in  $\Theta_0$  is thus the sequence  $\epsilon_n$  such that there exists a fixed positive constant  $C$  with

$$\limsup_{n \rightarrow \infty} \epsilon_n^{-1} R_n = C.$$

We say that a Bayesian procedure concentrates at the minimax rate if the concentration rate of the posterior in the class  $\Theta_0$  is the minimax convergence rate. Many models (prior and sampling models) studied in the literature have been proven to concentrate at the minimax rate in  $\Theta_0$ . In particular, in the density model, nonparametric mixture models are known to concentrate at the minimax rate (up to a log factor) over classes of Hölder functions for various types of kernels, see Ghosal and van der Vaart (2001), Ghosal and van der Vaart (2007), Shen et al. (2013) for Gaussian kernels, Kruijer et al. (2010) for location scale mixtures or Rousseau (2010) for beta kernels. Many other types of priors have been proven to lead to the minimax concentration rate, van der Vaart and Van Zanten (2008) proved minimax concentration rates of the posterior for Gaussian process priors, Ghosal and van der Vaart (2007) and Knapik et al. (2011) show minimax concentration rates for series expansions priors for regression and the white noise model respectively, Arbel et al. (2013), ?, Belitser and Ghosal (2003) obtained generic results for various sampling models.

The subspaces  $\Theta_0$  are restricted through regularity assumptions such as Sobolev or Hölder smoothness, shape restriction, or sparsity. These classes of functions are in general indexed by a parameter, say  $\beta$ , that accounts for the level of regularity or sparsity. In general the posterior concentration rate crucially depends

on this parameter. However, it is often difficult to fix  $\beta$  a priori, it is thus natural to seek procedures that perform well over a wide variety of  $\beta$  values, say  $\beta \in I$ . Such procedures are called *adaptive* as they automatically adapt the concentration rate over the whole collection of spaces  $\Theta_{0,\beta \in I}$ . Frequentist adaptive estimators have been well studied in the literature for the past three decades, see for instance Efroimovich (1986), Polyak and Tsybakov (1990), or Tsybakov (2009) for a review. From a Bayesian perspective, adaptive procedures have become more and more popular, see Belitser and Ghosal (2003) for infinite dimensional Gaussian distributions, Scricciolo (2006) obtained adaptive rates in the density model, van der Vaart and van Zanten (2009) considered Gaussian random fields priors, De Jonge et al. (2010) considered location scale mixtures. Other examples of adaptive Bayesian procedures can be found in Rivoirard et al. (2012), Rousseau (2010) or Arbel et al. (2013) for instance.

### 1.3 Nonparametric Bayesian testing

Another aspect of Bayesian nonparametric inference that has been investigated in this work is the so called testing problem or model choice. In this case, one is not interested in recovering an unknown parameter  $\theta$  but rather in taking a decision on the parameter given the observations. This problem of testing in a Bayesian framework is well known and can be dated back to Laplace (1814). It can be formalized as follows: let  $\Theta_0$  and  $\Theta_1$  be two distinct subspaces of the parameter space  $\Theta$ , associated with prior probability  $\pi_0$  and  $\pi_1$ , one wants to infer whether  $\theta \in \Theta_0$  versus  $\theta \in \Theta_1$ , which can be seen as the estimation of  $\mathbb{I}_{\Theta_1}(\theta)$  as argued in Robert (2007). Consider  $\Pi$  a prior distribution on  $\Theta = \Theta_0 \cup \Theta_1$ . In this setting it is natural to consider the 0-1 loss with weights  $\gamma_0, \gamma_1$  similar to the one proposed by Neyman and Pearson (1938) which is defined for a decision  $\varphi$

$$L(\theta, \varphi) = \begin{cases} \gamma_0 & \text{if } \varphi = 0 \text{ and } \mathbb{I}_{\Theta_0}(\theta) = 0 \\ \gamma_1 & \text{otherwise} \end{cases}.$$

The Bayesian solution to this problem (i.e. the minimizer of the Bayesian risk) is then

$$\varphi(\mathbf{X}^n) = \begin{cases} 1 & \text{if } \Pi(\Theta_1|\mathbf{X}^n)/\pi_1 \geq \frac{\gamma_0}{\gamma_0+\gamma_1} \Pi(\Theta_0|\mathbf{X}^n)/\pi_0 \\ 0 & \text{otherwise} \end{cases}. \quad (1.6)$$

To avoid the impact of  $\Pi(\Theta_0)$  and  $\Pi(\Theta_1)$  or  $\gamma_0$  and  $\gamma_1$ , one can equivalently define the Bayes-Factor

$$B_{0,1} = \frac{\Pi(\Theta_0|\mathbf{X}^n)}{\Pi(\Theta_1|\mathbf{X}^n)} \times \frac{\pi_1}{\pi_0}.$$

The testing procedure corresponds to rejecting  $\Theta_0$  if  $B_{0,1}$  is small but the Bayes-Factor provides more information than just a 0-1 answer. Standard thresholds are given by Jeffreys' scale. A test procedure based on the Bayes-Factor  $B_{0,1}$  is said to be consistent if  $B_{0,1}$  goes to infinity in  $P_{\theta_0}^n$  probability for all  $\theta_0 \in \Theta_0$  and converges to 0 in  $P_{\theta_0}^n$  probability for all  $\theta_0 \in \Theta_1$ . Bayes-Factors for nonparametric goodness of fit test have been studied in term of their asymptotic properties in the literature, see Dass and Lee (2004); McVinish et al. (2009); Rousseau (2007); Rousseau and Choi (2012) for instance. When both hypotheses are nonparametric and one is embedded in the other, the determination of Bayesian procedures that have good asymptotic properties is difficult in general.

Similarly to the estimation problem, asymptotic properties of a Bayesian answer to a testing problem are of great interest from both a theoretical and a methodological point of view since inference based on inconsistent posteriors could be highly misleading. A similar requirement should also hold for testing procedures. In this context, we will say that a procedure is consistent if it gives the right answer with probability that goes to 1 as the amount of information grows to infinity. More precisely, a testing procedure (1.6) is said to be consistent for the metric or semi-metric  $d$ , if for all  $\rho > 0$

$$\sup_{\theta \in \Theta_0} E_{\theta}(\varphi(\mathbf{X}^n)) = o(1), \quad \sup_{d(\theta, \Theta_0) > \rho} E_{\theta}(1 - \varphi(\mathbf{X}^n)) = o(1). \quad (1.7)$$

Similarly to the frequentist literature, we consider here *uniform* consistency, however this definition of consistency slightly differs from the one usually considered in the frequentist setting, as here we do not fix a level for the type I error of the test. It is also interesting to study the counterpart of the concentration rate in the testing problem namely the separation rate of the test. The separation rate is defined as the smallest sequence  $\rho = \rho_n$  such that (1.7) is still valid. It indicates how fast the test can differentiate both hypotheses. Similarly to the concentration rate, it also indicates which part of the prior influences the Bayesian procedure even asymptotically. This is of great interest as it is well known that in testing problems, the sensitivity to the prior is a major issue.

## 1.4 Challenging asymptotic properties of Bayesian nonparametric procedures

We have seen that studying the asymptotic behaviour of the posterior distribution is a major tool to understand the influence of the prior in the nonparametric setting. We have also seen that there exists sufficient conditions on the model under which the procedure is known to be consistent and to have optimal asymptotic behaviour. However, many statistical problems that are of interest in practice

do not fall under this general theory. These models present a new challenge for the Bayesian nonparametric community. In this section we present two of these problems namely the inference under monotonicity constraints and estimation in linear ill-posed inverse problems.

### 1.4.1 Inference under monotonicity constraints

In many statistical problems, it is useful to impose some restrictions on the parameter space to be able to carry out the inference. When modelling real world situations, shape constraints on the parameter of interest may appear naturally, this is the case for instance for drug response models or in survival analysis. Furthermore, these hypotheses are often easy to interpret, understand and explain compared to smoothness restrictions for instance. Among different shape constraints, monotonicity restrictions have been fairly popular in the literature. In a regression setting for instance, a monotonicity of a response is often granted from physical or theoretical considerations. Shape constraints inference, and monotonicity in particular can be dated back to Brunk (1955) and most of the early works on the subject can be found in Barlow et al. (1972). Since then monotonicity constraints have been used in many applied problems: in pharmaceutical context in Bornkamp and Ickstadt (2009), for survival analysis in Laslett (1982), Neelon and Dunson (2004) studied monotone regression for trend analysis and Dunson (2005) considered monotonicity constraints on count data. Many other applications can be found in Robertson et al. (1988).

In this section, we present the two shape constrained problems studied in this thesis, namely the estimation of a density under monotonicity constraints and testing for monotonicity in a regression setting.

#### 1.4.1.1 Monotone densities

Monotone densities are common in practice, especially in survival analysis. A first study of monotone density can be imputed to Grenander (1956) who considered the maximum likelihood estimator of a monotone density. Since then many others have been interested in estimating a unknown distribution under shape restrictions. Laslett (1982) considers the problem of estimating the distribution of cracks length on a mine wall, Sun and Woodroffe (1996) present some application in astronomy and renewal analysis among others. Using shape constraints procedures will ensure that the estimate follows this constraints, which could be a requirement of the analysis.

Since Williamson (1956), it is known that a density is monotone non increasing if and only if it is a mixture of uniform kernels. More precisely, let  $\mathcal{F}$  be the set of monotone non increasing densities on  $[0, \infty)$ , then for all  $f \in \mathcal{F}$  there exists a

probability distribution  $P$  such that

$$f(\cdot) = \int_0^\infty \frac{\mathbb{I}_{[0,\theta]}(\cdot)}{\theta} dP(\theta). \quad (1.8)$$

This mixture representation is particularly interesting as it allows for inference based on the likelihood. Grenander (1956) showed that the maximum likelihood estimator coincides with the first derivative of the least concave majorant of the cumulative distribution function. Its asymptotic properties were later studied in Groeneboom (1985) under the  $L_1$  loss and Prakasa Rao (1970) studied the asymptotic behaviour of the maximum likelihood estimator evaluated at a fixed point in the interior of the support. In Groeneboom (1989), it is shown that the minimax rate of convergence for this problem is of the order of  $n^{-1/3}$ . This shows in a way how monotonicity constraints act as regularity constraints as in this case, one obtains the same convergence rate as for Lipschitz densities. Another surprising aspect of monotone non increasing densities is that the evaluation of the maximum likelihood estimator at the boundaries of the support leads to inconsistent estimators. This problem has been studied in Sun and Woodroffe (1996) and very precise results on the behaviour of the maximum likelihood estimator at 0 can be found in Balabdaoui et al. (2011). More recently, Durot et al. (2012) obtain some asymptotic results for the maximum likelihood estimator under the supremum loss. In the Bayesian framework, monotone densities have been studied in Brunner and Lo (1989) and Lo (1984). From a Bayesian point of view, the mixture representation (1.8) leads naturally to a mixture type prior. Choosing a prior model on  $P$  in representation (1.8) naturally induces a prior on  $\mathcal{F}$ . This is the approach considered in Brunner and Lo (1989). In Chapter 2 we consider two types of priors on  $P$  namely Dirichlet process and finite mixtures with a random number of components. An interesting feature of these models is that the prior does not put positive mass on the Kullback-Leibler neighborhood of the truth, and thus condition (1.5) will not hold and the standard approach based on the work of Ghosal and van der Vaart (2007) cannot be applied directly. We prove that a similar result holds when one only considers Kullback-Leibler neighbourhoods of truncated versions of the densities

$$f_n(\cdot) = \frac{f(\cdot)\mathbb{I}_{[0,x_n]}}{F(x_n)}, \quad f_{\theta_0,n}(\cdot) = \frac{f_{\theta_0}(\cdot)\mathbb{I}_{[0,x_n]}}{F_{\theta_0}(x_n)},$$

where  $x_n$  is an increasing sequence and  $F$  is the cumulative distribution function of  $f$ . From this result, we prove that for both prior models, the posterior concentrates at the minimax rate  $n^{-1/3}$  up to a  $\log(n)$  term. We also study the asymptotic properties of the posterior distribution of the density at a fixed point  $x$  of its support. This is typically a difficult problem in general as Bayesian methods are in

general well suited for losses that are related to the Kullback-Leibler divergence (see Arbel et al., 2013; Hoffmann et al., 2013). In particular, the usual approach of Le Cam (1986) for constructing exponentially consistent sequence of tests does not hold in this case. However, we prove in Chapter 2 that for the considered prior distribution the posterior distribution of  $f(x)$  is consistent for every  $x$  in the support of  $f$ , including the boundaries. The fact that the posterior distribution is consistent at the boundaries of the support when the maximum likelihood estimator is not can be imputed to the penalization induced by the prior. Another interesting feature of our Bayesian approach is that the posterior is also consistent for the supremum loss over the whole support. Here again, the supremum loss is not related to the Kullback-Leibler divergence, which makes the construction of exponentially consistent sequence of tests difficult.

#### 1.4.1.2 Nonparametric test for monotonicity

Although there is a wide literature on the problem of estimating an unknown function under shape constraints, an important question is whether it is appropriate to impose a specific shape constraint. If it is, then the estimation procedures could in general be greatly improved by using a shape constrained estimation procedure. Conversely, imposing shape constraints in an appropriate case could lead to dramatically erroneous results. The problem of testing for monotonicity has been widely studied in the frequentist literature. Bowman et al. (1998) introduced a test for monotonicity in the regression setting base on the idea of *critical bandwidth* introduced in Silverman (1981). Hall and Heckman (2000) showed that this procedure is highly sensitive to flat parts of the regression function, and proposed another test procedure based on running gradient. Baraud et al. (2003), Ghosal et al. (2000b) and Baraud et al. (2005) propose testing procedures in the fixed design regression setting and the Gaussian white noise setting. Durot (2003) and Akakpo et al. (2014) consider a test that exploits the concavity of a primitive of a monotone function. A Bayesian approach to testing monotonicity in a regression framework has been proposed in Scott et al. (2013).

In Chapter 3, we consider the nonparametric regression model

$$Y_i = f(x_i) + \sigma\epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, 1), \quad (1.9)$$

and we want to test

$$H_0 : f \in \mathcal{F}, \text{ versus } H_1 : f \notin \mathcal{F}, \quad (1.10)$$

for  $\mathcal{F}$  be the set of monotone non increasing functions on  $[0, 1]$ .

A first difficulty in testing for monotonicity in a regression setting is that both the null and the alternative hypotheses are nonparametric. As a general rule when using posterior probabilities for hypothesis testing, it is important to take into

account the sensitivity to the prior distribution. This is true for parametric models but is critical for nonparametric ones as in that case, as stated before, the prior can still influence the posterior asymptotically. A second and probably more important difficulty is the fact that when testing for monotonicity in a regression setting, the null hypothesis is embedded in the alternative. This problem is common in *goodness of fit* tests where one is interested in testing  $f = f_0$  versus  $f \neq f_0$ . This has been investigated in Dass and Lee (2004), Ghosal et al. (2008) or McVinish et al. (2009) among others in the density setting, or Rousseau and Choi (2012) in the regression problem. In this case a main difficulty is that a parameter in the null model can also be approximated by a parameter in the alternative model. In fact it has been proved in Walker et al. (2004) that the Bayes-Factor will asymptotically support the model with prior that satisfies the Kullback-Leibler property, some additional conditions may be required when both priors do.

In the case of testing for monotonicity, it seems that for a natural choice of prior, namely piecewise constant functions with random number of bins, the Bayes-Factor is not consistent. We thus propose an alternative test that is asymptotically equivalent to testing for monotonicity using a similar idea as approximating a point null hypothesis by a shrinking interval (see Rousseau, 2007). Denote by  $\mathcal{F}$  the set of monotone non increasing functions with support  $[0, 1]$  and let  $\tilde{d}$  be a metric or a semi-metric. Consider the test

$$H_0^a : \tilde{d}(f, \mathcal{F}) \leq \tau \text{ versus } H_1^a : \tilde{d}(f, \mathcal{F}) \geq \tau \quad (1.11)$$

where  $\tilde{d}(f, \mathcal{F}) = \inf_{g \in \mathcal{F}} \tilde{d}(f, g)$  and  $\tau$  is a given threshold. If  $\tau$  decreases toward 0, both tests (1.10) and (1.11) are asymptotically equivalent. We propose a calibration of the threshold  $\tau$ , the Bayesian answer to the test (1.11) associated with the 0-1 loss is consistent for the initial problem of testing (1.10) and gives good results in practice compared to the frequentist procedures. Furthermore, for a specific choice of prior, the proposed Bayesian test is easy to implement which is a great advantage compared to the existing methods.

We also study the separation rate of the test which gives insights on the efficiency of the procedure. The adaptive minimax separation rates for testing monotonicity has been derived in Baraud et al. (2005) and Dümbgen et al. (2001) over Hölder alternatives. Under similar assumptions, we prove that our procedure achieves the minimax separation rate up to a  $\log(n)$  factor.

### 1.4.2 Ill posed linear inverse problems

Another general class of models that became popular for statistical modelling since the 1960's is the so called inverse problems. They appear naturally when one only has access to indirect observations of the parameter of interest. This is the

case in many fields of applications: medical imaging (computerized tomography), econometry (instrumental variables), radio astronomy (interferometry), astronomy (blurred images of Hubble telescope) or seismology among many others. In the statistical setting this is modelled by considering that the data arise from a probability distribution whose parameter has been transformed by a known operator  $K$  that acts on the parameter space. In most cases, we can assume that the transformation  $K$  does not induce additional noise in the observations. The sampling model is thus modified to

$$\mathbf{X}^n \sim P_{K\theta}^n, \theta \in \Theta. \quad (1.12)$$

If the operator  $K$  can be inverted and if its inverse is continuous, then the general theory applies and inference on  $\theta$  does not differ from the usual framework. However, in many cases, the inverse of the operator is not continuous. In this case the problem is called *ill-posed* with respect to Hadamard's definition, as in this case a small noise in the data will be greatly amplified in the inference on  $\theta$ . An interesting class of operators which covers many applications is the class of *linear operators on Hilbert Spaces*. It is usually assumed that the operator  $K$  is compact and injective and the Hilbert spaces are separable.

Statistical approach to inverse problems has grown popular since the standard framework has been proposed in Tikhonov (1963). A usual toy example to study such methods is the white noise model

$$\mathbf{X}^n = K\theta + \sigma \frac{W}{\sqrt{n}}, \quad (1.13)$$

where  $W$  is white noise and  $\sigma > 0$  a variance parameter. In this example we can easily grasp the difficulties at hand. In Chapter 4 we treat more general inverse problems models of the form (1.12). In the following, we will recall some features of statistical inference in inverse problems and illustrate it with model (1.13).

#### 1.4.2.1 Singular value decomposition

Consider  $K$  to be a compact injective linear operator between two Hilbert spaces  $\{\Theta, \langle \cdot, \cdot \rangle_\theta\}$  and  $\{\Xi, \langle \cdot, \cdot \rangle_\xi\}$ . For reading convenience, we shall drop the subscript for the inner product when there is no confusion. A usual approach to infer on  $\theta$  is to consider its decomposition in a basis of  $\Theta$ . In the linear inverse problem setting, a simple choice for such basis would be the one that diagonalize the operator  $K$ . More precisely, denote by  $K^*$  the adjoint operator of  $K$  and suppose that the auto-adjoint operator  $K^*K$  is compact, then the spectral Theorem states that  $K^*K$  has a complete orthogonal system of eigenvectors  $\{e_i\}$  with corresponding eigenvalues  $\{b_i\}$ . We thus have for all  $\theta \in \Theta$

$$K^*K\theta = \sum_{i=1}^{\infty} b_i \langle \theta, e_i \rangle e_i = \sum_{i=1}^{\infty} \kappa_i^2 \theta_i e_i, \quad (1.14)$$

where  $\kappa_i = \sqrt{b_i}$  and  $\theta_i = \langle \theta, e_i \rangle$ . In this case we say that  $K$  admits a singular value decomposition (SVD) with singular values  $\{\kappa_i\}$  and singular basis  $\{e_i\}$ . Inferring on  $\theta$  is thus equivalent to infer on the infinite sequence  $\{\theta_i\}$ . From the observations  $\mathbf{X}^n$  from model (1.12), one can get an estimator  $\hat{\eta}$  of  $\eta = K\theta$ . Denoting  $\{\hat{\eta}_i\}$  its projection onto the SVD basis, a simple estimator  $\hat{\theta}$  of  $\theta$  is given by

$$\hat{\theta}_i = \frac{\hat{\eta}_i}{\kappa_i}.$$

When the problem is ill-posed, since  $\{\kappa_i\}$ , goes to 0, we see that the coefficients  $\theta_i$  will be over-estimated for large  $i$ .

To see this problem more clearly, consider the white noise example. By projecting (1.13) onto the basis  $\{e_i\}$  and since  $W$  is a white noise, we can rewrite the model as

$$x_i = \kappa_i \theta_i + \frac{\sigma}{\sqrt{n}} \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1), \quad i = 1, 2, \dots$$

with  $x_i = \langle x, e_i \rangle$ . This sequence model has been a cornerstone in the study of linear inverse problems, see for instance Donoho (1995); Cavalier and Tsybakov (2002); Cavalier (2008). The case where  $K$  is the identity operator (i.e.  $\kappa_i = 1$  for all  $i$ ) has been widely studied in the literature. From a Bayesian perspective, this representation is highly interesting as in this case, it is natural to consider a prior on the sequence  $\{\theta_i\}$ . These types of priors have been considered in Ghosal and van der Vaart (2007) when  $K$  is the identity or Knapik et al. (2011) or Agapiou et al. (2013) in the inverse problem setting. To infer on  $\theta$ , we consider the transformed model

$$x_i \kappa_i^{-1} = \theta_i + \frac{\kappa_i^{-1} \sigma}{\sqrt{n}} \epsilon_i, \quad i = 1, 2, \dots,$$

which reduces the problem to estimating the mean of an infinite Gaussian sequence. Since the problem is ill-posed, the sequence  $\kappa_i^{-1} \rightarrow \infty$ , hence the variance of the noise blows up.

It appears from these considerations that the difficulty of an inverse problem can be quantified by the rate at which the sequence  $\{\kappa_i^{-1}\}$  goes to infinity.

**Definition 1.5** (Ill-posedness). We define the degree of ill-posedness of an inverse problem as follows:

- We say that a problem is *mildly ill-posed* of degree  $p$  if the sequence of singular values  $\{\kappa_i\}$  is such that there exist constants  $0 < C_d \leq C_u < \infty$  such that

$$C_d i^{-p} \leq \kappa_i \leq C_u i^{-p}.$$

- We say that a problem is *severely ill-posed* of degree  $p$  if the sequence of singular values  $\{\kappa_i\}$  is such that there exist constants  $0 < C_d \leq C_u < \infty$  and  $\gamma$  such that

$$C_d e^{-\gamma i^p} \leq \kappa_i \leq C_u e^{-\gamma i^p}.$$

Some generalized versions of the definition of ill-posedness of an operator have been considered in the literature (see Ray, 2013, for instance), however for the sake of simplicity, we will stick to this simple notion. The degree of ill-posedness greatly influence the complexity of a model. In particular, the minimax convergence rate for these models strongly depends on it, together with smoothness assumptions on  $\Theta$ .

#### 1.4.2.2 Examples of inverse problems

Even if for some operators the SVD is difficult to compute, and thus the degree of ill-posedness difficult to assess, there exists a series of classical operators for which the form of the SVD is explicit. Here we present some examples of ill-posed inverse problems that have been extensively studied in the literature.

**Numerical differentiation** If the problem of numerical integration has been well studied in practice and is well understood from a theoretical point of view, it turns out that the problem of numerical differentiation is much more complicated even for simple classes of functions. The operator  $K$  is thus defined for all  $\theta \in L_2([0, 1])$  by

$$K\theta(x) = \int_0^x \theta(u) du, \quad \forall x \in [0, 1].$$

The SVD is in this case given by the Fourier basis  $\{e_j\}$  and we easily obtain

$$K\theta = \sum_{j=-\infty}^{\infty} (2\pi i j)^{-1} \langle \theta, e_j \rangle e_j.$$

Thus the problem is mildly ill-posed of degree 1. We presented here the case of one time differentiation but similar results hold for the  $m$  time differentiation problem. It is mildly ill-posed of degree  $m$ .

**Deconvolution** A common problem in image processing is deconvolution of a signal. A particular example is image deblurring for instance. A standard framework is to consider circular deconvolution, that is for  $\theta$  and  $\lambda$  in  $L_2([0, 1])$  and 1-periodic, the operator  $K$  is defined as

$$K\theta(x) = \theta \star \lambda(x) = \int_0^1 \theta(u) \lambda(x - u) du, \quad \forall x \in [0, 1].$$

In this case, one only has access to a weighted average of  $f$  around the point  $x$ . Standard algebra gives that the singular basis is here again the Fourier basis and the singular values are the Fourier coefficients of the convolution kernel  $\lambda$ .

### 1.4.2.3 Regularization methods

As stated before the difficulty in inferring on the unknown parameter in inverse problems comes from the fact that the inverse of the operator  $K$  is not continuous over the all Hilbert space  $\Xi$ . A usual way to overcome this problem is to consider regularization methods to obtain a sensible estimator for these models. We present here two standard methods that are commonly used in practice. For a complete overview of regularization techniques, we refer to the monograph Engl et al. (1996).

Consider the general setting presented above, and consider a fixed sequence of weights  $w = \{w_i\}$  and an estimator  $\hat{\eta}$  of  $\eta = K\theta$ . Each sequence defines an estimator of  $\theta$

$$\hat{\theta}_i = w_i \frac{\hat{\eta}_i}{\kappa_i}, \quad \hat{\theta} = \sum_{i=1}^{\infty} \hat{\theta}_i e_i.$$

For a general sequence  $w$ , this estimator behaves poorly, due to the fact that for large  $i$ ,  $\kappa_i$  will be very small and will overwhelm the signal in  $\hat{\eta}_i$ . The simplest choice for the weight sequence  $w$  to bypass this problem is the projection sequence  $w_i = \mathbb{I}_{i \leq N}$  for some fixed threshold  $N$ . This regularization method is commonly called *spectral cut-off*. This calibration is rather rough as the weight only takes values 0 or 1, furthermore it requires a fine calibration of the bandwidth  $N$ .

Another approach is the celebrated *Tikhonov regularization* (Tikhonov and Arsenin, 1977) which is based on finding a minimizer of the data misfits while controlling the regularity of the estimator. The estimator is then obtained by

$$\hat{\theta} = \arg \min_{\theta} \{ \|K\theta - \mathbf{X}^n\|^2 + \mu \|\theta\|^2 \},$$

where  $\mu$  is a fixed tuning parameter. Here again the calibration of  $\mu$  is crucial. In particular, an optimal calibration in the minimax sense - i.e. that would lead to the minimax rate of convergence - will crucially depend on the regularity assumptions on  $\theta$  and the ill-posedness of the problem. If it is common to assume that the operator (and thus the degree of ill-posedness) is known, imposing a degree of regularity to the function  $\theta$  is a rather strong assumption. There exist data driven calibrations of  $\mu$  and  $N$ , however these are often difficult to study and will not be presented here.

### 1.4.2.4 The Bayesian approach to ill posed inverse problems

The Bayesian approach for ill-posed inverse problems is thus fairly natural as it is well known that putting a prior distribution on the unknown parameter often

acts as a regularization. This property is particularly useful in the model choice problem, but also for estimation as shown in ? in overfitted mixture models, or in Castillo (2013) when regularization is needed. Some of the priors proposed in the literature can be directly linked to the usual regularization methods. For instance the sieve prior presented in Ray (2013) corresponds to the spectral cut-off regularization. If the Bayesian approach to inverse problems has been put in practice (see for instance Orbanz and Buhmann, 2008), there is a dramatical lack of theoretical results for these models, and the families of prior distributions for which theoretical results exist are very limited.

Agapiou et al. (2013), Knapik et al. (2011) and Knapik et al. (2013) studied asymptotic properties of the posterior distribution for the conjugate (i.e. Gaussian) prior in the white noise setting. Minimax adaptive posterior concentration rates have been obtained in Knapik et al. (2012) also for conjugate priors. Ray (2013) considered a more general class of prior distributions that are still closely linked to the SVD of the operator. Moreover the general approach proposed by Ray (2013) leads to suboptimal rates in some cases. Thus it seems that there is a need for general results as the ones proposed in Ghosal and van der Vaart (2007) for the direct model.

In Chapter 4 we propose a general approach to derive posterior concentration rate for general ill-posed inverse problems. Our approach does not rely on a specific form of the prior distribution. With this result, we recover the known results in the literature and improve the suboptimal upper bounds for the posterior concentration rate obtained in Ray (2013). Furthermore, we derived posterior concentration rates for models that are neither conjugate nor related to the SVD of the operator. We consider an abstract setting in which the parameter space  $\mathcal{F}$  is an arbitrary metrizable topological vector space and let  $K$  be an injective mapping  $K : \mathcal{F} \ni f \mapsto Kf \in K\mathcal{F}$ . Even if the problem is ill-posed there exist subsets  $\mathcal{S}_n$  of  $K\mathcal{F}$  over which the inverse of the operator can be controlled. For suitably well chosen priors, these sets will capture most of the posterior mass, and we can thus easily derive posterior concentration rate for  $f$  from posterior concentration rate for  $Kf$  by a simple inversion of the operator.

A main contribution of this thesis is to study the asymptotic behaviour of the posterior distributions for problems for which general results do not hold. In Chapter 2 we study the problem of estimating monotone non increasing densities. In Chapter 3 we focus on the problem of testing monotonicity of a regression function. Finally in Chapter 4 we provide general conditions to derive posterior concentration rates for ill-posed linear inverse problems. Many other models presented in the literature may require such non standard methods to study the asymptotic behaviour of the posterior distribution.

## Bibliography

- Agapiou, S., Larsson, S., and Stuart, A. M. (2013). Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Process. Appl.*, 123(10):3828–3860.
- Akakpo, N., Balabdaoui, F., and Durot, C. (2014). Testing monotonicity via local least concave majorants. *Bernoulli*, 20(2):514–544.
- Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, 2:1152–1174.
- Arbel, J., Gayraud, G., and Rousseau, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scandinavian Journal of Statistics*, 40(3):549–570.
- Balabdaoui, F., Jankowski, H., Pavlides, M., Seregin, A., and Wellner, J. (2011). On the Grenander estimator at zero. *Statist. Sinica*, 21(2):873–899.
- Baraud, Y., Huet, S., and Laurent, B. (2003). Adaptive tests of qualitative hypotheses. *ESAIM Probab. Stat.*, 7:147–159.
- Baraud, Y., Huet, S., and Laurent, B. (2005). Testing convex hypotheses on the mean of a Gaussian vector. Application to testing qualitative hypotheses on a regression function. *Ann. Statist.*, 33(1):214–257.
- Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression*. John Wiley & Sons, London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics.
- Belitser, E. and Ghosal, S. (2003). Adaptive bayesian inference on the mean of an infinite-dimensional normal distribution. *Annals of statistics*, pages 536–559.
- Bornkamp, B. and Ickstadt, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose-response analysis. *Biometrics*, 65(1):198–205.
- Bowman, A., Jones, M., and Gijbels, I. (1998). Testing monotonicity of regression. *Journal of computational and Graphical Statistics*, 7(4):489–500.
- Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, 26(4):607–616.
- Brunner, L. J. and Lo, A. Y. (1989). Bayes methods for a symmetric unimodal density and its mode. *Ann. Statist.*, 17(4):1550–1566.
- Castillo, I. (2013). On bayesian supremum norm contraction rates. *arXiv preprint arXiv:1304.1761*.
- Cavalier, L. (2008). Nonparametric statistical inverse problems. *Inverse Problems*, 24(3):034004, 19.
- Cavalier, L. and Tsybakov, A. (2002). Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields*, 123(3):323–354.
- Dass, S. C. and Lee, J. (2004). A note on the consistency of bayes factors for testing point null versus non-parametric alternatives. *Journal of statistical planning and*

- inference*, 119(1):143–152.
- De Jonge, R., Van Zanten, J., et al. (2010). Adaptive nonparametric bayesian inference using location-scale mixture priors. *The Annals of Statistics*, 38(6):3300–3320.
- Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.*, 14(1):1–67. With a discussion and a rejoinder by the authors.
- Donoho, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.*, 2(2):101–126.
- Doob, J. L. (1949). Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, Colloques Internationaux du Centre National de la Recherche Scientifique, no. 13, pages 23–27. Centre National de la Recherche Scientifique, Paris.
- Dümbgen, L., Spokoiny, V. G., et al. (2001). Multiscale testing of qualitative hypotheses. *The Annals of Statistics*, 29(1):124–152.
- Dunson, D. B. (2005). Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association*, 100(470):618–627.
- Durot, C. (2003). A kolmogorov-type test for monotonicity of regression. *Statistics & probability letters*, 63(4):425–433.
- Durot, C., Kulikov, V. N., Lopuhaä, H. P., et al. (2012). The limit distribution of the  $l_{\infty}$ -error of grenander-type estimators. *The Annals of Statistics*, 40(3):1578–1608.
- Efroimovich, S. (1986). Nonparametric estimation of a density of unknown smoothness. *Theory of Probability & Its Applications*, 30(3):557–568.
- Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230.
- Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.*, 2:615–629.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian data analysis*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, second edition.
- Ghosal, S., Ghosh, J., and van der Vaart, A. (2000a). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.
- Ghosal, S., Lember, J., Van Der Vaart, A., et al. (2008). Nonparametric bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89.
- Ghosal, S., Sen, A., and van der Vaart, A. W. (2000b). Testing monotonicity of regression. *Ann. Statist.*, 28(4):1054–1082.
- Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distribu-

- tions for non-i.i.d. observations. *Ann. Statist.*, 35(1):192–223.
- Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5):1233–1263.
- Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian nonparametrics*. Springer Series in Statistics. Springer-Verlag, New York.
- Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.*, 39:125–153 (1957).
- Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 539–555, Belmont, CA. Wadsworth.
- Groeneboom, P. (1989). Brownian motion with a parabolic drift and Airy functions. *Probab. Theory Related Fields*, 81(1):79–109.
- Hall, P. and Heckman, N. E. (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *Ann. Statist.*, 28(1):20–39.
- Hoffmann, M., Rousseau, J., and Schmidt-Hieber, J. (2013). On adaptive posterior concentration rates. *arXiv preprint arXiv:1305.5270*.
- Knapik, B., van Der Vaart, A., Van Zanten, J., et al. (2011). Bayesian inverse problems with gaussian priors. *The Annals of Statistics*, 39(5):2626–2657.
- Knapik, B. T., Szabó, B. T., van der Vaart, A. W., and van Zanten, J. H. (2012). Bayes procedures for adaptive inference in inverse problems for the white noise model. *ArXiv e-prints*.
- Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. (2013). Bayesian recovery of the initial condition for the heat equation. *Comm. Statist. Theory Methods*, 42(7):1294–1313.
- Kruijer, W., Rousseau, J., and van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.*, 4:1225–1257.
- Laplace, P. (1814). *Essai philosophique sur les probabilités*; Mme. Ve. Courcier.
- Laslett, G. (1982). The survival curve under monotone density constraints with applications to two-dimensional line segment processes. *Biometrika*, 69(1):153–160.
- Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York.
- Le Cam, L. and Yang, G. L. (2000). *Asymptotics in statistics*. Springer Series in Statistics. Springer-Verlag, New York, second edition. Some basic concepts.
- Lenk, P. J. (1991). Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, 78(3):531–543.
- Lijoi, A., Prünster, I., and Walker, S. G. (2007). Bayesian consistency for stationary models. *Econometric Theory*, 23(4):749–759.

- Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.*, 12(1):351–357.
- McVinish, R., Rousseau, J., and Mengersen, K. (2009). Bayesian goodness of fit testing with mixtures of triangular distributions. *Scandinavian journal of statistics*, 36(2):337–354.
- Müller, P. and Mitra, R. (2013). Bayesian nonparametric inference – why and how. *Bayesian Analysis*, 8(2):269–302.
- Neelon, B. and Dunson, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics*, 60(2):398–406.
- Neyman, J. and Pearson, E. S. (1938). *Contributions to the theory of testing statistical hypotheses*. University Press.
- Orbanz, P. and Buhmann, J. M. (2008). Nonparametric bayesian image segmentation. *International Journal of Computer Vision*, 77(1-3):25–45.
- Polyak, B. T. and Tsybakov, A. B. (1990). Asymptotic optimality of the  $c_p$ -test for the orthogonal series estimation of regression. *Theory of Probability & Its Applications*, 35(2):293–306.
- Prakasa Rao, B. L. S. (1970). Estimation for distributions with monotone failure rate. *Ann. Math. Statist.*, 41:507–519.
- Ray, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electron. J. Stat.*, 7:2516–2549.
- Rivoirard, V., Rousseau, J., et al. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Analysis*, 7(2):311–334.
- Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer.
- Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester.
- Rousseau, J. (2007). Approximating interval hypothesis: p-values and bayes factors. In J.M., B., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., editors, *Bayesian statistics 8: proceedings of the eighth Valencia International Meeting, June 2-6, 2006*, number vol. 8 in Oxford science publications. Oxford University Press.
- Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.*, 38(1):146–180.
- Rousseau, J. and Choi, T. (2012). Bayes factor consistency in non iid models. Technical report, Technical report.
- Schwartz, L. (1965). On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 4:10–26.
- Scott, J. G., Shively, T. S., and Walker, S. G. (2013). Nonparametric Bayesian

- testing for monotonicity. *ArXiv e-prints*.
- Scribciolo, C. (2006). Convergence rates for bayesian density estimation of infinite-dimensional exponential families. *The Annals of Statistics*, 34(6):2897–2920.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica*, 4(2):639–650.
- Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive bayesian multivariate density estimation with dirichlet mixtures. *Biometrika*, 100(3):623–640.
- Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Annals of Statistics*, pages 687–714.
- Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 97–99.
- Sun, J. and Woodroffe, M. (1996). Adaptive smoothing for a penalized NPMLE of a non-increasing density. *J. Statist. Plann. Inference*, 52(2):143–159.
- Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*, 101(476):1566–1581.
- Tikhonov, A. N. (1963). On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk SSSR*, 151:501–504.
- Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of ill-posed problems*. V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York-Toronto, Ont.-London. Translated from the Russian, Preface by translation editor Fritz John, Scripta Series in Mathematics.
- 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.
- van der Vaart, A. W. and Van Zanten, J. (2008). Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, pages 1435–1463.
- van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, pages 2655–2675.
- Walker, S. (2003). On sufficient conditions for Bayesian consistency. *Biometrika*, 90(2):482–488.
- Walker, S. (2004). New approaches to Bayesian consistency. *Ann. Statist.*, 32(5):2028–2043.
- Walker, S., Damien, P., and Lenk, P. (2004). On priors with a kullback–leibler property. *Journal of the American Statistical Association*, 99(466).
- Walker, S. and Hjort, N. L. (2001). On Bayesian consistency. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 63(4):811–821.
- Williamson, R. E. (1956). Multiply monotone functions and their Laplace trans-

forms. *Duke Math. J.*, 23:189–207.

# Chapter 2

## Monotone densities

“Now we are joined together and have been since noon. And no one to help either of us.”

– **Ernest Hemingway**, *The old man and the sea*.

### Résumé

Dans ce chapitre, nous étudions la consistance et la vitesse de concentration de la loi a posteriori dans le modèle de densité décroissante pour différentes métriques. Ce modèle est particulièrement intéressant car les densités décroissantes ont une représentation sous forme de mélange d’uniformes et sont donc un cas particulier de mélange pour lequel le support du noyau dépend du paramètre. Dans ce cadre, les hypothèses classiques nécessaires pour la consistance de la loi a posteriori ne sont pas vérifiées. Notamment la loi a priori ne met pas suffisamment de masse sur les voisinages de Kullback-Leibler du vrai paramètre, et une adaptation des méthodes usuelles est donc nécessaire. Pour deux familles d’a priori classiques, nous prouvons que l’a posteriori se concentre à la vitesse maximax pour les pertes  $L_1$  et Hellinger. Nous étudions ensuite la consistance de la loi a posteriori de la densité pour les pertes ponctuelle et norme sup. Ces deux métriques sont en général difficiles à étudier car elles ne peuvent être reliées à la divergence naturelle qu’est la divergence de Kullback-Leibler. Pour ces deux pertes, nous prouvons la consistance de l’a posteriori et donnons une borne supérieure pour la vitesse de concentration.

## 2.1 Introduction

The nonparametric problem of estimating monotone curves, and monotone densities in particular, has been well studied in the literature both from a theoretical and applied perspectives. Shape constrained estimation is fairly popular in the nonparametric literature and widely used in practice (see Robertson et al., 1988, for instance). Monotone densities appear in a wide variety of applications such as survival analysis, where it is natural to assume that the uncensored survival time has a monotone non increasing density. In these problems, estimating the survival function is equivalent to estimate the survival time density say  $f$  and the pointwise estimate  $f(0)$ . It is thus interesting to have a better understanding of the behaviour of the estimation procedures in this case. An interesting property of monotone non increasing densities on  $\mathbb{R}^+$  is that they have a mixture representation pointed out by Williamson (1956)

$$f(x) = \int_0^\infty \frac{\mathbb{I}_{[0,\theta]}(x)}{\theta} dP(\theta), \quad (2.1)$$

where  $P$  is a probability distribution on  $\mathbb{R}^+$  called the mixing distribution. In order to emphasize the dependence in  $P$ , we will denote  $f_P$  the functions admitting representation (2.1). This representation allows for inference based on the likelihood. Grenander (1956) derived the nonparametric maximum likelihood estimator of a monotone density and Prakasa Rao (1970) studied the behavior of the Grenander estimator at a fixed point. Groeneboom (1985) and more recently, Balabdaoui and Wellner (2007) studied very precisely the asymptotic properties of the non parametric maximum likelihood estimator. It is proved to be consistent and to converge at the minimax rate  $n^{-1/3}$  when the support of the distribution is compact. In their paper Durot et al. (2012) get some refined asymptotic results for the supremum norm.

The mixture representation of monotone densities lead naturally to a mixture type prior on the set of monotone non increasing densities with support on  $[0, L]$  or  $\mathbb{R}^+$ . For example Ferguson (1983) and Lo (1984) introduced the Dirichlet Process prior (DP) and Brunner and Lo (1989) considered the special case of unimodal densities with a prior based on a Dirichlet Process mixture. The problem of deriving concentration rates for mixtures models have receive a huge interest in the past decade. Wu and Ghosal (2008) studied properties of general mixture models Ghosal and van der Vaart (2001) studied the well known problem of Gaussian mixtures, Rousseau (2010) derive concentration rates for mixtures of betas, Kruijer et al. (2010) proved good adaptive properties of mixtures of Gaussian. Extensions to the multivariate case have recently been introduced (e.g. Shen et al.

(2013)).

Under monotonicity constrained, we derive an upper bound for the posterior concentration rate with respect to some metric or semi metric  $d(\cdot, \cdot)$ , that is a positive sequence  $(\epsilon_n)_n$  that goes to 0 when  $n$  goes to infinity such that

$$E_0^n (\Pi(d(f, f_0) > \epsilon_n | \mathbf{X}^n)) \rightarrow 0,$$

where the expectation is taken under the true distribution  $P_0$  of the data  $\mathbf{X}^n$  and where  $f_0$  is the density of  $P_0$  with respect to the Lebesgue measure. Following Khazaei et al. (2010) we study two families of nonparametric priors on the class of monotone non increasing densities. Interestingly in our setting, the so called Kullback-Leibler property, that is the fact that the prior puts enough mass on Kulback-Leibler neighbourhood of the true density, is not satisfied. Thus the approach based on the seminal paper of Ghosal et al. (2000) cannot be applied. We therefore use a modified version of their results and obtain for the two families of prior a concentration rate of order  $(n/\log(n))^{-1/3}$  which is the minimax estimation rate up to a  $\log(n)$  factor under the  $L_1$  or Hellinger distance. We extend these results to densities with support on  $\mathbb{R}^+$  and prove that under some conditions on the tail of the distribution, the posterior still concentrates at an almost optimal rate. To the author's knowledge, no concentration rates have been derived for monotone densities on  $\mathbb{R}^+$ .

Interestingly, the non parametric maximum likelihood estimator of  $f_P(x)$  is not consistent for  $x = 0$  (see Sun and Woodroffe (1996) and Balabdaoui and Wellner (2007) for instance). However, we prove that the posterior distribution of  $f$  is still consistent at this point under a specific family of non parametric mixture prior. In fact we prove the pointwise consistency of the posterior for all  $x$  in  $[0, L]$  with  $L \leq \infty$ . We then derive a consistent Bayesian estimator of the density at any fixed point of the support. This is particularly interesting as the point-wise loss is usually difficult to study in a Bayesian framework as the Bayesian approaches are well suited to losses related to the Kullback-Leiber divergence. We also study the behaviour of the posterior distribution for the sup norm when the density has a compact support. This problem has been addressed recently in the frequentist literature by Durot et al. (2012). They derive refined asymptotic results on the sup norm of the difference between a Grenander-type estimator and the true density on sub intervals of the form  $[\epsilon, L - \epsilon]$  where  $\epsilon > 0$  avoiding the problems at the boundaries. Here, we prove that the posterior distribution is consistent in sup norm on the whole support of  $f_0$  when it has compact support. We also derive concentration rate for the posterior of the density taken at a fixed point and for the sup norm on subsets of  $[0, L]$  for  $L < \infty$ . We also derive an upper bound for the concentration rate of  $f(x)$  for  $x \in (0, L)$  but only get suboptimal rates using

a testing approach as in Giné and Nickl (2010). It is to be noted that for this problem the modulus of continuity for the pointwise and Hellinger losses defined for  $f_0 \in \mathcal{F}$  and  $x \in (0, L)$  by

$$m(\epsilon) := \sup\{|f(x) - f_0(x)| : f \in \mathcal{F}, h(f, f_0) \leq \epsilon\}$$

is of the order  $\epsilon^{2/3}$  (see Donoho and Liu, 1991). Given the discussion in Hoffmann et al. (2013), it is to be expected that the usual approach of Ghosal et al. (2000) based on tests will lead to suboptimal concentration rates. We now introduce some notations which will be needed throughout the paper.

**Notations** For  $0 < L \leq \infty$  define the set  $\mathcal{F}_L$  by

$$\mathcal{F}_L = \left\{ f \text{ s.t. } 0 \leq f < \infty, f \searrow, \int_0^L f = 1 \right\},$$

We also define  $\mathfrak{S}_k$  the  $k$ -simplex that is the set  $\{(s_1, \dots, s_k) \in [0, 1]^k, \sum_{i=1}^k s_i = 1\}$ . Let  $KL(p_1, p_2)$  be the Kullback Leibler deviation between the densities  $p_1$  and  $p_2$  with respect to some measure  $\lambda$

$$KL(p_1, p_2) = \int \log \left( \frac{p_1}{p_2} \right) p_1 d\lambda.$$

We also define the Hellinger distance  $h(p_1, p_2)$  between  $p_1$  and  $p_2$  as

$$h^2(p_1, p_2) = \frac{1}{2} \int (\sqrt{p_1} - \sqrt{p_2})^2 d\lambda.$$

We will say that  $\Xi^n = o_{p_0}(1)$  if  $\Xi^n \rightarrow 0$  under  $P_0$ . Finally we will denote  $f'$  the derivative of  $f$ .

**Construction of a prior distribution on  $\mathcal{F}_L$**  Using the mixture representation of monotone non increasing densities (2.1) we construct nonparametric priors on the set  $\mathcal{F}_L$  by considering a prior on the mixing distribution  $P$ . Let  $\mathcal{P}$  be the set of probability measures on  $[0, L]$ . Thus we fall in the well known set up of nonparametric mixture priors models. We consider two types of prior on the set  $\mathcal{P}$ .

**Type 1 : Dirichlet Process prior**  $P \sim DP(A, \alpha)$  where  $A$  is a positive constant and  $\alpha$  a probability density on  $[0, L]$ .

**Type 2 : Finite mixture**  $P = \sum_{j=1}^K p_j \delta_{x_j}$  with  $K$  a non zero integer and  $\delta_x$  the dirac function on  $x$ . We choose a prior distribution  $Q$  on  $K$  and given  $K$ , define distributions  $\pi_{x,K}$  on  $(x_1, \dots, x_K) \in [0, L]^K$  and  $\pi_{p,K}$  on  $(p_1, \dots, p_K) \in \mathfrak{S}_K$ .

For  $\mathbf{X}^n = (X_1, \dots, X_n)$ , a sample of  $n$  independent and identically distributed random variables with common probability distribution function  $f$  in  $\mathcal{F}_L$  with respect to the Lebesgue measure, we denote  $\Pi(\cdot | \mathbf{X}^n)$  the posterior probability measure associated with the prior  $\Pi$ .

The paper is organised as follow: the main results are given in Section 2.2, where conditions on the priors are discussed. The proofs are presented in Section 2.3.

## 2.2 Main results

Concentration rates of the posterior distributions have been well studied in the literature and some general results link the rate to the prior (see Ghosal et al. (2000)). However, in our setting, the Kullback Leibler property is not satisfied in its usual form and thus the standard Theorems do not hold. In fact an interesting feature of mixture distributions whose kernels have varying support is that the prior mass of the sets  $\{f, KL(f_0, f) = +\infty\}$  is 1 for most  $f_0 \in \mathcal{F}_L$  given that  $f$  and  $f_0$  will have different support. One could prevent this by imposing that the support of the mixing distribution is wider than the support of  $f_0$ , however this could lead to a deterioration of the concentration rate. Here, we use a modified version of the results of Ghosal et al. (2000) considering truncated versions of the density  $f$ . This idea has been considered in Khazaei et al. (2010) in a similar setting. We impose some conditions on the prior under which the posterior distribution concentrates at the minimax rate up to a  $\log(n)$  term.

### Conditions on the prior

**C1 condition on  $\alpha$**  Let  $\alpha$  be a probability density on  $\mathbb{R}^+$  such that for all  $\theta \in (0, L)$ ,  $\alpha(\theta) > 0$ . Consider the following conditions on  $\alpha$

- for  $0 < t_1 \leq t_2$  and  $\theta$  small enough

$$\theta^{t_1} \lesssim \alpha(\theta) \lesssim \theta^{t_2} \quad (2.2a)$$

- for  $1 < a_1 \leq a_2$  and  $\theta$  small enough

$$e^{-a_1/\theta} \lesssim \alpha(\theta) \lesssim e^{-a_2/\theta} \quad (2.2b)$$

- for  $1 < b_1 \leq b_2$  and  $\theta$  small enough

$$e^{-b_1/\theta} \lesssim \alpha(L - \theta) \lesssim e^{-b_2/\theta} \quad (2.2c)$$

**C2 condition for Type I prior** For  $P \sim DP(\alpha, M)$  with  $\alpha$  satisfying C1

**C3 condition for the Type II prior** The following conditions holds

- For some positive constants  $C_1, C_2, a_1, \dots, a_k, c$

$$e^{-C_1 K \log(K)} \geq Q(K) \geq e^{-C_2 K \log(K)} \quad (2.3)$$

$$\pi_{p,k}(p_1, \dots, p_K) \geq K^{-K} c^K p_1^{a_1} \dots p_K^{a_K} \quad (2.4)$$

- $\pi_{x,K}$  is the distribution of  $K$  independent and identically distributed random variables sampled from  $\alpha$ .

**C4 Condition for densities on  $\mathbb{R}^+$**  If  $f_0 \in \mathcal{F}_\infty$  then for  $\beta$  and  $\tau$  some fixed positive constant we have for  $x$  large enough

$$f_0(x) \leq e^{-\beta x^\tau}. \quad (2.5)$$

### 2.2.1 Posterior concentration rate for the $L_1$ and Hellinger metric

The following Theorems gives the posterior concentration rate for the  $L_1$  and Hellinger metric for monotone non increasing densities on  $[0, L]$  with  $L < \infty$  and  $L = \infty$ . For both Theorems the proofs are postponed to section 2.3.

**Theorem 2.1.** *Let  $\mathbf{X}^n = (X_1, \dots, X_n)$  be an independent and identically distributed sample with a common probability distribution function  $f_0$  such that  $f_0 \in \mathcal{F}_L$  with  $0 < L < \infty$ . Let  $\Pi$  be either a Type I or Type II prior satisfying **C2** or **C3** respectively with  $\alpha$  satisfying (2.2a). If  $d(\cdot, \cdot)$  is either the  $L^1$  or Hellinger distance, then there exists a positive constant  $C$  such that*

$$\Pi \left( f, d(f, f_0) \geq C \left( \frac{n}{\log(n)} \right)^{-1/3} \mid \mathbf{X}^n \right) \rightarrow 0, \quad P_0 \text{ a.e.} \quad (2.6)$$

when  $n$  goes to infinity, where  $C$  depends on  $f_0$  only through  $L$  and an upper bound on  $f_0(0)$ . Furthermore, if for  $\delta > 0$ ,  $\sup_{[0, \delta]} |f'_0(x)| < \infty$  and  $\alpha$  satisfies (2.2b), or  $\sup_{[L, L-\delta]} |f'_0(x)| < \infty$  and  $\alpha$  satisfies (2.2c), then (2.6) still holds.

Conditions C1 and C2 are roughly the same as in Khazaei et al. (2010). Theorem 2.1 is thus an extension of their results to concentration rates. We also extend their results to mixtures prior satisfying (2.2b) or (2.2c) under some additional conditions on  $f_0$ . This will prove useful for the estimation of  $f_0$  and  $f_L$ . Under condition C3 on the tail of the true density, i.e. we require exponential tails, we get the posterior concentration rate for density with support on  $\mathbb{R}^+$ .

**Theorem 2.2.** *Let  $\mathbf{X}^n = (X_1, \dots, X_n)$  be an independent and identically distributed sample with a common probability distribution density  $f_0$  such that  $f_0 \in \mathcal{F}_\infty$  and  $f_0$  satisfy **C3**. Let  $\Pi$  be either a Type I or Type II prior satisfying **C2** or **C3** respectively with  $\alpha$  satisfying (2.2a). Then for some positive constant  $C$  we have for  $d(\cdot, \cdot)$  either the  $L_1$  or Hellinger metric*

$$\Pi \left( d(f_P, f_0) \geq C (n/\log(n))^{-1/3} \log(n)^{1/\tau} | \mathbf{X}^n \right) \rightarrow 0, \quad P_0 \text{ a.e.} \quad (2.7)$$

*when  $n$  goes to infinity. Similarly, if for  $\delta > 0$ ,  $\sup_{[0, \delta]} |f'_0(x)| < \infty$  and  $\alpha$  satisfies (2.2b), (2.7) still holds.*

Note that considering monotone non increasing densities on  $\mathbb{R}^+$  deteriorates the upper bound on the posterior concentration rate with a factor  $\log(n)^{1/\tau}$ . It is not clear whether it could be sharpen or not. For instance, in the frequentist literature, Reynaud-Bouret et al. (2011) observe a slower convergence rate when considering infinite support for densities without any other conditions. In a Bayesian setting, a similar log term appears in Kruijer et al. (2010) when considering densities with non compact support. However this deterioration of the concentration rate does not have a great influence on the asymptotic behaviour of the posterior. Note also that the tail conditions are mild since  $\tau$  can be taken as small as needed, and thus the considered densities can have almost polynomial tails.

The above results on the posterior concentration rate in terms of the  $L_1$  or Hellinger metric are new to our knowledge but not surprising. The specificity of these results lies in the fact that the usual approach based on the approach of Ghosal et al. (2000) need to bound the prior mass of Kullback Leibler neighbourhoods of the true density which cannot be done here as explained in section 2.1.

### 2.2.2 Consistency and posterior concentration rate for the pointwise and supremum loss

The following results consider the pointwise loss function for which only a few exist in the Bayesian nonparametric literature, see for instance the paper of Giné and Nickl (2010). The following Theorem proves consistency of the posterior distribution for all point in the interior of the support.

**Theorem 2.3.** *Let  $x$  be in  $(0, L)$  with  $0 < L \leq \infty$  but  $x < \infty$ . Let  $f_0 \in \mathcal{F}_L$  such that  $f'_0$  exists near  $x$  and  $f'_0(x) < 0$ . Let  $X_i$ ,  $i = 1, \dots, n$  and  $\Pi$  be either a Type I or Type II prior satisfying **C2** or **C3** respectively with  $\alpha$  satisfying **C1** with either (2.2a), (2.2b) or (2.2c). Then, for all  $x$  in  $(0, L)$  with  $x < \infty$ , and  $\epsilon > 0$*

$$\Pi(|f_P(x) - f_0(x)| > \epsilon | \mathbf{X}^n) \rightarrow 0. \quad (2.8)$$

Consider the posterior median  $\hat{f}_n^\pi(x) = \inf\{t, \Pi[f_P(x) \leq t | \mathbf{X}^n] > 1/2\}$ , it follows that

$$P_0(|\hat{f}_n^\pi(x) - f_0(x)| > \epsilon | \mathbf{X}^n) \rightarrow 0. \quad (2.9)$$

We thus have a pointwise consistency of the posterior distribution of  $f_0(x)$  for every  $x$  in the interior of the support of  $f_0$ . The maximum likelihood is not consistent at the boundaries of the support as pointed out in Sun and Woodroffe (1996) for instance. In particular it is not consistent at 0 and when  $L < \infty$ , it is not consistent at  $L$ . It is known that integrating the parameter as done in Bayesian approaches induces a penalisation. This is particularly useful in testing or model choice problems but can also be effective in estimation problems, see for instance Rousseau and Mengersen (2011). Here we require that the base measure puts exponentially small mass at the boundaries. This induce enough penalization to achieve consistency of the posterior distribution of  $f(0)$  and  $f(L)$ . The following Theorem gives consistency of the posterior distribution of  $f$  at every point on the support of  $f_0$  including the boundaries.

**Theorem 2.4.** *Let  $x$  be in  $[0, L]$  with  $0 < L \leq \infty$  but  $x < \infty$ . Let  $f_0 \in \mathcal{F}_L$  such that  $f'_0$  exists at  $x$  and  $f'_0(x) < 0$ . Let  $X_i$ ,  $i = 1, \dots, n$  and  $\Pi$  be either a Type I or Type II prior satisfying **C2** or **C3** with  $\alpha$  satisfying condition (2.2b) if  $x = 0$  or (2.2c) if  $x = L$ . Then, for all  $x$  in  $[0, L]$  with  $x < \infty$ , and  $\epsilon > 0$*

$$\Pi(|f_P(x) - f_0(x)| > \epsilon | \mathbf{X}^n) \rightarrow 0. \quad (2.10)$$

Consider the posterior median  $\hat{f}_n^\pi(x) = \inf\{t, \Pi[f_P(x) \leq t | \mathbf{X}^n] > 1/2\}$ , it follows that

$$P_0(|\hat{f}_n^\pi(x) - f_0(x)| > \epsilon | \mathbf{X}^n) \rightarrow 0. \quad (2.11)$$

The problem of estimating  $f_0(0)$  under monotonicity constraints is another example of the effectiveness of penalisation induced by integration on the parameters. Although we do not have a proof for inconsistency of the posterior of  $f(0)$  or  $f(L)$  when  $\alpha$  satisfies (2.2a), we believe that the similarly to the maximum likelihood estimator, the posterior distribution is in this case not consistent.

The following Theorem gives an upper bound on the concentration rate of the posterior distribution under the pointwise loss.

**Theorem 2.5.** *Let  $f_0$  be in  $\mathcal{F}_L$  with  $0 < L \leq \infty$  and  $\Pi$  be either a Type I or Type II prior satisfying **C1** or **C2** respectively with  $\alpha$  satisfying **C1**, and let  $x$  be in  $(0, L)$  such that  $f'$  exists in a neighbourhood of  $x$  and  $f'(x) < 0$ , then for  $C$  a positive constant*

$$\Pi\left(|f_P(x) - f_0(x)| > C \left(\frac{n}{\log(n)}\right)^{-2/9} | \mathbf{X}^n\right) \rightarrow 0. \quad (2.12)$$

when  $n$  goes to infinity.

Here the concentration rate is suboptimal. It is however the best rate that one can obtain using the usual approach by testing (see Hoffmann et al., 2013). Proving that the posterior concentrates at the rate  $n^{-1/3}$  up to some power of  $\log(n)$  would require some more refined control of the posterior distribution close to Bernstein von Mises types of results, see Castillo (2013), which in the case of mixture models is very difficult and beyond the scope of this chapter.

We derive from Theorem 2.4 the consistency of the posterior distribution for the sup norm. This is particularly useful when considering confidence bands, as pointed out in Giné and Nickl (2010). Under similar assumptions as in Durot et al. (2012), we get the consistency of the posterior distribution for the sup norm. Note that contrariwise to Durot et al. (2012), we do not restrict to sub-intervals of the support of the density. This is mainly due to the fact that the Bayesian approaches are consistent at the boundaries of the support of  $f_0$ .

**Theorem 2.6.** *Let  $f_0 \in \mathcal{F}_L$  with  $0 < L < \infty$  be such that  $f'_0$  exists and  $\|f'_0\|_\infty < \infty$  and for all  $x \in [0, L]$ ,  $f'_0(x) < 0$ . Let also the prior  $\Pi$  be either a Type I or Type II prior satisfying **C1** or **C2** with  $\alpha$  satisfying conditions (2.2b) and (2.2c) respectively. Then*

$$\Pi\left(\sup_{x \in [0, L]} |f_P(x) - f_0(x)| > \epsilon | X_n\right) \rightarrow 0. \quad (2.13)$$

Similar results as in Theorem 2.5 also hold for the concentration rate of the posterior distribution for the supremum over all subsets of the form  $(a, b)$  with  $0 < a < b < L$  with the same rate.

## 2.3 Proofs

In this section we prove Theorems 2.1 to 2.6 given in Section 2.2. To prove Theorems 3-6, we need to construct tests that are adapted to the pointwise or supremum loss. The usual approach based on ? cannot be applied in this case. We thus construct test based on the Maximum Likelihood Estimator.

### 2.3.1 Proof of Theorems 2.1 and 2.2

The proofs of Theorems 2.1 and 2.2 follow the general ideas of Ghosal et al. (2000) with some modification due to the fact that the Kullback-Leibler property is not satisfied. We first focus on density on  $\mathcal{F}_L$  with  $L < \infty$  and extend these results to monotone non increasing density with support  $\mathbb{R}^+$  that satisfy C3. We extended the approach used in Khazaei et al. (2010) to the concentration rate framework

and get similar results as those presented in Ghosal et al. (2000). More precisely, the proofs relies on the following Theorem which is a modification of Ghosal et al. (2000) main Theorem proposed by Rivoirard et al. (2012). To tackle the fact that the usual Kullback Leibler property is not satisfied in its usual sense, we consider truncated versions of the densities

$$f_n(\cdot) = \frac{f(\cdot)\mathbb{I}_{[0,\theta_n]}(\cdot)}{F(\theta_n)}, \quad f_{0,n}(\cdot) = \frac{f_0(\cdot)\mathbb{I}_{[0,\theta_n]}(\cdot)}{F_0(\theta_n)} \quad (2.14)$$

where  $\theta_n$  is defined as

$$\theta_n = \inf\{x, 1 - F_0(x) < \frac{\epsilon_n}{2n}\}.$$

We then define the counterpart of the Kullback Leibler neighbourhoods

$$S_n(\epsilon_n, \theta_n) = \left\{ f, KL(f_n, f_{0,n}) \leq \epsilon_n^2, \right. \\ \left. \int f_{0,n}(x) \left( \log \left( \frac{f(x)}{f_0(x)} \right) \right)^2 dx \leq \epsilon_n^2, \int_0^{\theta_n} f(x) dx \gtrsim 1 - \epsilon_n^2 \right\}. \quad (2.15)$$

**Theorem 2.7.** *Let  $f_0$  be the true density and let  $\Pi$  be a prior on  $\mathcal{F}$  satisfying the following conditions : there exist a sequence  $(\epsilon_n)$  such that  $\epsilon_n \rightarrow 0$  and  $n\epsilon_n^2 \rightarrow \infty$  and a constant  $c > 0$  such that for any  $n$  there exist  $\mathcal{F}_n \subset \mathcal{F}$  satisfying*

$$\Pi(\mathcal{F}_n^c) = o(\exp(-(c+2)n\epsilon_n^2)).$$

*For any  $j \in \mathbb{N}$ ,  $j > 0$ , let  $\mathcal{F}_{n,j} = \{f \in \mathcal{F}_n, j\epsilon_n < d(f, f_0) \leq (j+1)\epsilon_n\}$  and  $N_{n,j}$  the Hellinger (or  $L_1$ ) metric entropy of  $\mathcal{F}_{n,j}$ . There exists a  $J_{0,n}$  such that for all  $j \geq J_{0,n}$*

$$N_{n,j} \leq (K-1)n\epsilon_n^2 j^2,$$

*where  $K$  is an absolute constant.*

*Let  $S_n(\epsilon_n, \theta_n)$  be defined as in (2.15) and let  $\Pi$  be such that*

$$\Pi(S_n(\epsilon_n, \theta_n)) \geq \exp(-cn\epsilon_n^2). \quad (2.16)$$

*We have :*

$$\Pi(f : d(f_0, f) \leq J_{0,n}\epsilon_n | \mathbf{X}^n) = 1 + o_P(1).$$

The proof of this Theorem is postponed to Appendix 2.5. We will thus prove that the conditions of Theorem 2.7 are satisfied in our case. Let  $f_0$  be in  $\mathcal{F}_L$ . The following lemma states that (2.16) is satisfied.

**Lemma 2.1.** *Let  $\Pi$  be either a Type 1 or Type 2 prior on  $\mathcal{F}_L$  as in Theorem 2.1 and let  $S_n(\epsilon_n, \theta_n)$  be a set as in (2.15), then*

$$\Pi(S_n(\epsilon_n, \theta_n)) \gtrsim \exp \left\{ C_1 \epsilon_n^{-1} \log(\epsilon_n) \right\}. \quad (2.17)$$

This lemma is proved in appendix 2.4. The  $\epsilon$  metric entropy of the set of bounded monotone non increasing densities has been shown to be less than  $\epsilon^{-1}$ , up to a constant (see Groeneboom (1986) or van der Vaart and Wellner (1996) for instance). As the prior puts mass on  $\mathcal{F}_L$ , on which  $f(0)$  is not uniformly bounded, we consider an increasing sequence of sieves

$$\mathcal{F}_n = \{f \in \mathcal{F}_L, f(0) \leq M_n\}. \quad (2.18)$$

where  $M_n = \exp \left\{ cn^{1/3} \log(n)^{2/3} (t_2 + 1)^{-1} \right\}$  with  $t_2$  as in the conditions C1 or C2. The following Lemma shows that  $\mathcal{F}_n$  covers most of the support of  $\Pi$  as  $n$  increase.

**Lemma 2.2.** *Let  $\mathcal{F}_n$  be defined by (2.18) and  $\Pi$  be either a Type 1 or Type 2 as in Theorem 2.1, then*

$$\Pi(\mathcal{F}_n^c) \lesssim e^{-cn^{1/3} \log(n)^{2/3}}.$$

Here again, the proof is postponed to appendix 2.4. We now get an upper bound for the  $\epsilon$ -metric entropy of the set  $\mathcal{F}_n$ . Recall that in Groeneboom (1985) it is proved that the  $L_1$  metric entropy of monotone non increasing densities on  $[0, 1]$  bounded by  $M$  can be bounded from above by  $C_0 \log(M) \epsilon_n^{-1}$ . We cannot apply this result directly for the sets  $\mathcal{F}_n$  as it would give a suboptimal control of the entropy to construct tests in a similar way as in Ghosal et al. (2000). In fact the upper bound on the entropy of  $\mathcal{F}_n$  is of the order of  $e^{n\epsilon_n}$  the usual conditions of Ghosal et al. (2000) requires an upper bound of the order  $e^{n\epsilon_n^2}$ . However as stated in Theorem 2.7 it is enough to bound the  $\epsilon$ -metric entropy of the sets

$$\mathcal{F}_{n,j} = \{f \in \mathcal{F}_n, j\epsilon_n \leq d(f, f_0) \leq (j+1)\epsilon_n\},$$

for  $j \in \mathbb{N}^*$ . We can easily adapt the results of Groeneboom (1985) to positive monotone non increasing functions on any interval  $[a, b]$  and get the following Lemma.

**Lemma 2.3.** *Let  $\tilde{\mathcal{F}}$  be the set of positive monotone non increasing functions on  $[a, b]$  such that for all  $f$  in  $\tilde{\mathcal{F}}$ ,  $\int_a^b f \leq M_2$  and  $f \leq M$ , then*

$$N(\epsilon, \tilde{\mathcal{F}}, d) \lesssim \epsilon^{-1} \log(M+1) \left( (b-a) + 3M_2 \right).$$

The proof of this Lemma is straightforward given the results of Groeneboom (1985) and is thus omitted. Let  $x_{n,j} \in [0, L]$  such that  $\epsilon_n/2 \leq x_{n,j} \leq \epsilon_n$ . We denote for all  $f$  in  $\mathcal{F}_{n,j}$   $f_{1,j} = f\mathbb{I}_{[0, x_{n,j})}$  and  $f_{2,j} = f\mathbb{I}_{[x_{n,j}, L]}$ . Since for all  $f$  in  $\mathcal{F}_{n,j}$  we have  $\int_0^1 |f(x) - f_0(x)|dx \leq (j+1)\epsilon_n$  then

$$\int_0^{x_{n,j}} f(x)dx - \int_0^{x_{n,j}} f_0(x)dx \leq (j+1)\epsilon_n,$$

which implies that

$$x_{n,j}f(x_{n,j}) \leq x_{n,j}f_0(0) + (j+1)\epsilon_n,$$

which in turn gives

$$f(x_{n,j}) \leq f_0(0) + 2(j+1).$$

Recall that for all  $f \in \mathcal{F}_n$  we have  $f(0) \leq M_n$ . Using Lemma 2.3, we construct an  $\epsilon_n/2$ -net for the set  $\mathcal{F}_{n,j}^1 = \{f_{1,j}, f \in \mathcal{F}_{n,j}\}$  with  $N_1$  points, and

$$\log(N_1) \lesssim \epsilon_n^{-1} \log(M_n + 1)\epsilon_n(j+2),$$

and thus deduce

$$\log(N_1) \leq C'n\epsilon_n^2 j^2 \quad (2.19)$$

Similarly, given that  $f(x_{n,j}) \leq M + 2(j+1)$  we get an  $\epsilon_n/2$ -net for the set  $\mathcal{F}_{n,j}^2 = \{f_{2,j}, f \in \mathcal{F}_{n,j}\}$  with  $N_2$  points and

$$\log(N_2) \leq \tilde{C}'n\epsilon_n^2 j^2. \quad (2.20)$$

This provide a  $\epsilon_n$ -net for  $\mathcal{F}_{n,j}$  with less than  $N_1 \times N_2$  points. Given (2.19) and (2.20) the  $L_1$  metric entropy of the sets  $\mathcal{F}_{n,j}$  satisfy

$$\log(N(\mathcal{F}_{n,j}, \epsilon_n, L_1)) \lesssim n\epsilon_n^2 j^2. \quad (2.21)$$

The conditions of Theorem 2.7 are thus satisfied which ends the proof of Theorem 2.1

**Extention to  $\mathbb{R}^+$**  Given that  $f_0(x) \lesssim e^{-\beta x^\tau}$  when  $x$  goes to infinity, if  $\theta_n$  is such that  $\theta_n = \inf\{x, 1 - F_0(x) < \epsilon_n/(2n)\}$  then  $\theta_n \lesssim (\log(n))^{1/\tau}$ . Using similar arguments as before, Lemma 2.1 still holds under the exponential tail assumption. We now get an upper bound for the  $\epsilon$ -metric entropy of  $\mathcal{F}_{n,j}$ . Here again, we split  $\mathcal{F}_{n,j}$  into two parts. The construction of an  $\epsilon_n/2$ -net for  $\mathcal{F}_{n,j}^1$  does not change and therefore (2.19) holds. Finally, let  $\tilde{\mathcal{F}}_{n,j}^2 = \{f \in \mathcal{F}_{n,j}^2, \forall x > \theta_n, f(x) = 0\}$ . Given Lemma 2.3, we get for  $c_1 > 0$  large enough an  $\epsilon_n/(2c_1(j+1))$ -net for  $\tilde{\mathcal{F}}_{n,j}^2$  by considering  $f^*$  the restriction of  $f$  to  $[x_{n,j}, \theta_n]$ . We have

$$d(f, f^*) \leq c_2(j+1)\epsilon_n,$$

where  $d(\cdot, \cdot)$  is either the  $L_1$  or Hellinger distance. Hence, for  $c_1 > c_2$  an  $\epsilon/2$ -net for  $\mathcal{F}_{n,j}^2$  with at most  $e^{c_3 n \epsilon_n^2 j^2}$  points and thus

$$\log(N(\mathcal{F}_{n,j}^2, \epsilon_n, d)) \leq \tilde{C}'' n \epsilon_n^2 j^2.$$

We conclude using the same arguments as in the preceding section, and thus Theorem 2.2 is proved.

### 2.3.2 Proof of Theorems 2.3 and 2.5

To prove Theorem 2.3 and 2.5, we need to construct tests for all  $x \in (0, L)$  of  $f_0$  versus  $|f_P(x) - f_0(x)| \geq \epsilon_n^{2/3}$  as the approach used in Ghosal et al. (2000) is not suited for the pointwise loss. As we have  $\Pi(\|f_P - f_0\|_1 > \epsilon_n | \mathbf{X}^n) = o_{P_0}(1)$  we can consider functions  $f_P$  such that  $\|f_P - f_0\|_1 \leq \epsilon_n$ . We construct tests  $\Phi_n$  such that

$$E_0^n(\Phi) = o(1), \quad \sup_{f, |f(x) - f_0(x)| > \epsilon_n} E_f^n(1 - \Phi) \leq e^{-Cn\epsilon_n^2}.$$

Denote  $A_\epsilon^x := \{f, |f(x) - f_0(x)| > \epsilon\}$  that can be split into  $A_\epsilon^{x,+} = \{f, f(x) - f_0(x) > \epsilon\}$  and  $A_\epsilon^{x,-} = \{f, f(x) - f_0(x) < -\epsilon\}$  and denote  $e_n = e_0 \epsilon_n^{2/3}$  and  $h_n = h_0 e_n$ . Consider the tests

$$\begin{aligned} \phi_n^+ &= \mathbb{I} \left\{ n^{-1} \sum_{i=1}^n \mathbb{I}_{[x-h_n, x]}(X_i) - \int_{x-h_n}^x f_0(t) dt > c_n \right\} \\ \phi_n^- &= \mathbb{I} \left\{ n^{-1} \sum_{i=1}^n \mathbb{I}_{[x, x+h_n]}(X_i) - \int_x^{x+h_n} f_0(t) dt < -c_n \right\} \end{aligned}$$

We immediately get  $E_0^n(\max(\phi_n^+, \phi_n^-)) = o(1)$ . Note that if  $f_P(x) > f_0(x) + e_n$  then

$$\begin{aligned} \int_{x-h_n}^x f_P(t) - f_0(t) dt &\geq h_n(f_P(x) - f_0(x)) - \int_{x-h_n}^x f_0(t) - f_0(x) dt \\ &\geq h_n e_n - C_0 h^2 \end{aligned}$$

for some  $C_0 > 0$  that only depends on  $f_0$ . Similarly if  $f_P(x) < f_0(x) - e_n$  then for all  $h > 0$

$$\int_x^{x+h} f_P(t) - f_0(t) dt \leq -h e_n + C_0 h^2$$

We thus deduce for  $f_P$  such that  $f_P(x) - f_0(x) > e_n$

$$\begin{aligned} P_f(1 - \phi_n^+) &\leq P_f \left( n^{-1} \sum_{i=1}^n \mathbb{I}_{[x-h_n, x]}(X_i) - \int_{x-h_n}^x f_P(t) dt \leq -h_n e_n + C_0 h^2 + c_n \right) \\ &\leq P_f \left( n^{-1} \sum_{i=1}^n \mathbb{I}_{[x-h, x]}(X_i) - \int_{x-h}^x f_P(t) dt \leq -h_0 e_n^2 / 2 \right), \end{aligned}$$

if  $c_n \leq e_n^2$  and  $h_0 \leq 1/C_0$ . Now note that for  $f_P$  such that  $\|f_P - f_0\|_1 \leq \epsilon_n$

$$\begin{aligned} \int_{x-h_n}^x f_P &\geq - \int_0^\infty |f - f_0| + \int_{x-h_n}^x f_0 \\ &\geq -\epsilon_n + \int_{x-h_n}^x f_0 \\ &\geq -e_n + h_n f_0(x) \geq h_n f_0(x) / 2. \end{aligned}$$

Moreover,

$$\int_{x-h_n}^x f_P \leq e_n + h_n f_0(x - h_n) \leq 2h_n f_0(x)$$

for  $n$  large enough and  $h$  small enough. We conclude that

$$\text{Var}_{f_P}^n \left( n^{-1} \sum_{i=1}^n \mathbb{I}_{[x-h, x]}(X_i) \right) \leq 2h f_0(x)$$

Thus using Bernstein's inequality (e.g. van der Vaart and Wellner (1996) Lemma 2.2.9 p. 102) we get

$$P_f(1 - \phi^+) \leq 2e^{-nh_n e_n^2 / (2 + e_n/3)}.$$

Similarly, we have

$$P_f(1 - \phi_n^-) \leq 2e^{-nh_n e_n^2 / (2 + e_n/3)}.$$

Taking  $\Phi_n = \max(\phi_n^+, \phi_n^-)$  we deduce

$$\begin{aligned} P_0(\Phi_n) &= o(1) \\ \sup_{f \in A_{e_n}^x} P_f(1 - \Phi_n) &\leq e^{-Ch_0 e_n^3} \end{aligned}$$

We have

$$P_0(\Phi_n) = o(1)$$

$$\sup_{f \in A_{e_n}^x} P_f(1 - \Phi_n) \leq e^{-Cne_0\epsilon_n^2}$$

Similarly to the proof of Theorem 2.7, following Khazaei et al. (2010), we get an exponentially small lower bound for  $D_n$ . More precisely, we get that

$$D_n \geq 2e^{-(c+2)n\epsilon_n^2}$$

with probability that goes to 1. Note that

$$\begin{aligned} \mathbf{E}_0^n \left( \frac{N_n}{D_n} \right) &\leq \mathbf{E}_0^n(\Phi_n^x) + P_0^n(D_n \leq e^{-(c+2)n\epsilon_n^2}) + \\ &\quad \mathbf{E}_0^n(\Pi[\mathcal{F}_n^c | \mathbf{X}^n]) + e^{(c+2)n\epsilon_n^2} \int_{A_\epsilon \cap \mathcal{F}_n} \mathbf{E}_f^n(1 - \Phi_n^x) d\Pi(f) \end{aligned} \quad (2.22)$$

Given the preceding results, we have

$$\mathbf{E}_0^n \left( \frac{N_n}{D_n} \right) \leq o(1) + e^{(c+2)n\epsilon_n^2} \sup_f \mathbf{E}_f^n(1 - \Phi_n^x)$$

which ends the proof choosing  $e_0$  large enough.

**Consistency of a Bayesian estimator** We consider in this section  $\hat{f}_n^\pi(t)$ , the Bayesian estimator associated with the absolute error loss, define as the median of the posterior distribution. Consistency of the posterior mean, which is the most common Bayesian estimator is however not proved here but could nevertheless be an interesting result.

We first define  $\hat{f}_n^\pi(t)$  such that

$$\hat{f}_n^\pi(t) = \inf\{x, \Pi[f_P(t) \leq x | \mathbf{X}^n] > 1/2\}. \quad (2.23)$$

In order to get consistency in probability we note that if  $\hat{f}_n^\pi(t) - f_0(t) > \epsilon$  then

$$\Pi(f_P(t) > f_0(t) + \epsilon | \mathbf{X}^n) > 1/2.$$

And if  $\hat{f}_n^\pi(t) - f_0(t) < -\epsilon$  then

$$\Pi(f_P(t) < f_0(t) - \epsilon | \mathbf{X}^n) > 1/2.$$

We deduce, with Markov inequality and Theorem 2.3

$$\begin{aligned}
P_0^n(\hat{f}_n^\pi(t) - f_0(t) > \epsilon) &\leq P_0^n(\Pi(f_P(t) > f_0(t) + \epsilon | \mathbf{X}^n) > 1/2) \\
&\leq 2\mathbf{E}_0^n(\Pi(f_P(t) > f_0(t) + \epsilon | \mathbf{X}^n) > 1/2) \\
&\leq o(1),
\end{aligned}$$

and similarly

$$P_0^n(\hat{f}_n^\pi(t) - f_0(t) < -\epsilon) \leq o(1).$$

Thus we have  $P_0^n(|\hat{f}_n^\pi(t) - f_0(t)| > \epsilon) \rightarrow 0$  which gives the consistency in probability of  $\hat{f}_n^\pi(t)$ .

### 2.3.3 Proof of Theorem 2.4

The previous proof holds for all  $x \in (0, L)$  we now need to prove the consistency of the posterior for  $x = 0$  and  $x = L$ , when the prior satisfies conditions (2.2b) or (2.2c). We first consider the case  $x = 0$ , the case  $x = L$  can be deduce with symmetric arguments.

As before, consider the set  $A_\epsilon^0$  and split it in  $A_\epsilon^{0,+}$  and  $A_\epsilon^{0,-}$ . Note that using the same test  $\phi_n^-$  as before we easily get

$$\Pi(A_\epsilon^{0,-} | \mathbf{X}^n) = o_{P_0}(1).$$

We now consider  $f_P \in A_\epsilon^{0,+}$ . As before we can restrict ourselves to functions  $f_P$  such that  $\|f_P - f_0\|_1 \leq \epsilon_n$ . We thus have for  $h = 2\epsilon_n/\epsilon$

$$\begin{aligned}
f_P(0) - f_0(0) &\leq f_P(0) - f_P(h) + h^{-1} \int |f_0(t) - f_P(t)| dt \\
&\leq f_P(0) - f_P(h) + h^{-1} \epsilon_n \\
&= f_P(0) - f_P(h) + \epsilon/2.
\end{aligned}$$

We now prove that the prior mass of the event  $\{f_P(0) - f_P(h) > \epsilon/2\}$  is less than  $e^{-(c+2)n\epsilon_n^2}$ . Using Markov inequality we get

$$\Pi(f_P(0) - f_P(h) > \epsilon/2) \leq 2\epsilon^{-1} \int_0^h \frac{1}{\theta} \alpha(\theta) d\theta \leq e^{-a_2/h} \lesssim e^{-a_2 n \epsilon_n^2 \log(n)}.$$

Using the same control for  $D_n$  as in the proof of Theorem 2.7, and applying the usual method of Ghosal et al. (2000), we get the desired result.

### 2.3.4 Proof of Theorem 2.6

In this section we prove that the posterior distribution is consistent in sup norm. Here again, the main difficulty is to construct tests that are adapted to the considered loss. More precisely we construct a test  $\Phi$  such that

$$E_0^n(\Phi) = o(1), \quad \sup_{f, \sup_{[0,L]} |f-f_0| > \epsilon_n} E_f^n(1 - \Phi) \leq e^{-Cn\epsilon_n^2}.$$

To do so we consider a combination of the tests considered in the previous section noting that if the posterior distribution is consistent at the points of a sufficiently refined partition of  $[0, L]$  then it is consistent for the sup norm. Here again, we will only consider the case  $L = 1$  without loss of generality. We first denote

$$B_\epsilon = \left\{ f, \sup_{[0,L]} |f(x) - f_0(x)| > \epsilon \right\}$$

Let  $C'_0$  be a positive constant such that  $\|f'_0\|_\infty \leq C'_0$  and let  $(x_i)_i$  be the separation points of a  $\epsilon/(8C'_0)$  regular partition of  $[0, 1]$  and  $p = \text{Card}\{(x_i)_i\}$ . Note that

$$B_\epsilon = \bigcup_{i=1}^p \{f, \sup_{[x_i, x_{i+1}]} |f(x) - f_0(x)| > \epsilon\}.$$

Recall that  $A_\epsilon^x = \{f, |f(x) - f_0(x)| > \epsilon\}$ . We consider the set  $B_\epsilon \cap \bigcap_{i=1}^p (A_{\epsilon/8}^{x_i})^c$ . Given Theorem 2.3, we have that

$$E_0^n \left( \Pi \left( \bigcup_{i=1}^p (A_{\epsilon/5}^{x_i}) \middle| \mathbf{X}^n \right) \right) = o(1).$$

If  $f \in B_\epsilon$  we have for all  $x \in [x_i, x_{i+1}]$ ,

$$|f(x) - f_0(x)| \leq |f(x) - f(x_i)| + |f(x_i) - f_0(x_i)| + |f_0(x_i) - f_0(x)|.$$

Given that  $f$  is monotone non increasing, and given the hypotheses on  $f_0$  we have

$$\begin{aligned} |f(x) - f(x_i)| &\leq |f(x_{i+1}) - f(x_i)| \\ &\leq |f(x_{i+1}) - f_0(x_{i+1})| + |f_0(x_{i+1}) - f_0(x_i)| + |f_0(x_i) - f(x_i)| \\ &\leq 3\epsilon/5 \end{aligned}$$

and for the same reasons

$$|f(x_i) - f_0(x_i)| + |f_0(x_i) - f_0(x)| \leq 2\epsilon/5.$$

Which leads to

$$|f(x) - f_0(x)| \leq \epsilon$$

and thus, taking the supremum over  $x$ , we get

$$\sup_{x \in [x_i, x_{i+1}]} |f(x) - f_0(x)| \leq \epsilon.$$

We then deduce

$$\Pi(B_\epsilon | \mathbf{X}^n) \leq \Pi \left( B_\epsilon \cap \left\{ \bigcap_{i=1}^p (A_{\epsilon/5}^{x_i})^c \right\} \right) + \Pi \left( \bigcup_{i=1}^p (A_{\epsilon/5}^{x_i}) \right) = o_{P_0}(1)$$

Which gives the consistency of the posterior distribution in sup norm

## 2.4 Technical Lemmas

### 2.4.1 Proof of Lemma 2.1

To prove Lemma 2.1, we first construct stepwise constant functions such that these approximations are in the truncated Kullback Leibler neighbourhood of  $f_0$ . We then construct a set  $\mathcal{N}$  included in  $S_n(\epsilon_n, \theta_n)$  based on the considered piecewise constant approximation such that for  $\Pi$  a Type I or Type II prior  $\Pi(\mathcal{N}) \geq e^{-Cn\epsilon_n^2}$ .

We first construct a piecewise constant approximation of  $f_0$  which is base on a sequential subdivision of the interval  $[0, L]$  with more refined subdivisions where  $f_0$  is less regular such that the number of points is less than  $\epsilon_n^{-1}$  points.

This approximation is adapted from the proof of Theorem 2.5.7 in van der Vaart and Wellner (1996). We then identify a finite piecewise constant density by a mixture of uniform for which the Hellinger distance between the piecewise constant approximation  $f_P$  of  $f_0 \in \mathcal{F}$  and  $f_0$  is less than  $\epsilon_n$  and  $\|f_0/f_P\|_\infty \leq M$ . The following Lemma gives the form of a finite probability distribution  $P$  such that  $f_P$  is in the Kullback-Leibler neighbourhood of some  $f \in \mathcal{F}$ .

**Lemma 2.4.** *Let  $f \in \mathcal{F}_L$  be such that  $f(0) \leq M < +\infty$ . For all  $0 < \epsilon < 1$  there exists  $m \lesssim L^{1/3} M^{1/3} \epsilon^{-1}$ ,  $p = (p_1, \dots, p_m) \in \mathfrak{S}_m$  and  $x = (x_1, \dots, x_m) \in [0, L]^m$  such that  $P = \sum_{i=1}^m \delta_{x_i} p_i$  satisfies*

$$KL(f, f_P) \lesssim \epsilon^2, \quad \int \left( \log \left( \frac{f}{f_P} \right) \right)^2 f \lesssim \epsilon^2, \quad (2.24)$$

where  $f_P$  is defined as in (2.1).

*Proof.* For a fixed  $\epsilon$ , let  $f$  be in  $\mathcal{F}_L$ . Consider  $\mathcal{P}_0$  the coarsest partition :

$$0 = x_0^0 < x_1^0 = L,$$

at the  $i^{th}$  step, let  $\mathcal{P}_i$  be the partition

$$0 = x_0^i < x_1^i < \dots < x_{n_i}^i = L,$$

and define

$$\varepsilon_i = \max_j \{ (f(x_{j-1}^i) - f(x_j^i))(x_j^i - x_{j-1}^i)^{1/2} \}.$$

For each  $j \geq 1$ , if  $(f(x_{j-1}^i) - f(x_j^i))(x_j^i - x_{j-1}^i)^{1/2} \geq \frac{\varepsilon_i}{\sqrt{2}}$  we split the interval  $[x_{j-1}^i, x_j^i]$  into two subsets of equal length. We then get a new partition  $\mathcal{P}_{i+1}$ . We continue the partitioning until the first  $k$  such that  $\varepsilon_k^2 \leq \epsilon^3$ . At each step  $i$ , let  $n_i$  be the number of intervals in  $\mathcal{P}_i$ ,  $s_i$  the number of interval in  $\mathcal{P}_i$  that have been divided to obtain  $\mathcal{P}_{i+1}$ , and  $c = 1/\sqrt{2}$ . Thus, it is clear that  $\varepsilon_{i+1} \leq c\varepsilon_i$

$$\begin{aligned} s_i(c\varepsilon_i)^{2/3} &\leq \sum_j (f(x_{j-1}^i) - f(x_j^i))^{2/3} (x_j^i - x_{j-1}^i)^{1/3} \\ &\leq \left( \sum_j f(x_{j-1}^i) - f(x_j^i) \right)^{2/3} \left( \sum_j x_j^i - x_{j-1}^i \right)^{1/3} \leq M^{2/3} L^{1/3}, \end{aligned}$$

using Hölder inequality. We then deduce that

$$\begin{aligned} \sum_{j=1}^k n_j &= k + \sum_{j=1}^k j s_{k-j} \leq 2 \sum_{j=1}^k j s_{k-j} \leq 2 \sum_{j=1}^k j M^{2/3} L^{1/3} (c\varepsilon_{k-j})^{-2/3} \\ &\leq 2M^{2/3} L^{1/3} \varepsilon_k^{-2/3} 2^{1/3} \sum_{j=1}^k j 2^{-j/3} \\ &\leq K_0 M^{2/3} L^{1/3} \varepsilon_k^{-2/3}, \end{aligned}$$

where  $K_0 = 2(1 - 2^{-2/3})^{-2}$ . Thus

$$n_k \leq K_0 M^{2/3} L^{1/3} \epsilon^{-1}. \quad (2.25)$$

Now, for  $f \in \mathcal{F}_L$ , we prove that there exists a stepwise density with less than  $K_0 M^{2/3} L^{1/3} \frac{1}{\epsilon}$  pieces such that

$$KL(f, h) \leq \epsilon^2 \text{ and } \int f \log\left(\frac{f_0}{f_P}\right)^2(x) dx \lesssim \epsilon^2 \quad (2.26)$$

In order to simplify notations, we define

$$x_i = x_i^k, \quad l_i = x_i - x_{i-1}, \quad g_i = f(x_{i-1})^{1/2}.$$

We consider the partition constructed above associated with  $f^{1/2}$ , which is also a monotone nonincreasing function that satisfy  $f^{1/2}(0) \leq M^{1/2}$  (instead of  $M$ ). We denote  $g$  the function defined as  $g(x) = \sum \mathbb{I}_{[x_{i-1}, x_i]}(x) g_i$

$$\begin{aligned} \|f^{1/2} - g\|_2^2 &= \int (f^{1/2} - g)^2(x) dx = \sum_{i=1}^{n_k} \int_{I_i} (f^{1/2} - g)^2(x) dx \\ &\leq \sum_{i=1}^{n_k} \int_{I_i} (f^{1/2}(x_{i-1}^k) - f^{1/2}(x_i^k))^2 dx \\ &\leq \sum_{i=1}^{n_k} (x_i^k - x_{i-1}^k) (f^{1/2}(x_{i-1}^k) - f^{1/2}(x_i^k))^2 \\ &\leq n_k \varepsilon_k^2 \leq L^{1/3} K_0 M^{1/3} \epsilon^2. \end{aligned}$$

We then define  $h = \frac{g^2}{\int g^2}$  and get an equivalent of  $\int g^2$ .

$$\begin{aligned} \int g^2 dx &= \int (g^2 - f)(x) dx + 1 \\ &= \int (g - \sqrt{f})(g + \sqrt{f})(x) dx + 1 \\ &= 1 + \mathcal{O}(\epsilon), \end{aligned}$$

and deduce that  $(\int g^2)^{1/2} = 1 + \mathcal{O}(\epsilon)$ . Let  $H$  be the Hellinger distance

$$\begin{aligned} H(f, h) &= H\left(f, \frac{g^2}{\int g^2}\right) \\ &\leq H(f, g^2) + H(g^2, \frac{g^2}{\int g^2}) \\ &\leq L^{1/6} K_0 M^{1/6} \epsilon + \left(\int (g - \frac{g}{(\int g^2)^{1/2}})^2(x) dx\right)^{1/2} \lesssim \epsilon. \end{aligned}$$

Since  $\|f/h\|_\infty = \|f/g^2\|_\infty (\int g^2) \leq (\int g^2)$ , together with the above bound on  $H(f, h)$  and Lemma 8 from Ghosal and van der Vaart (2007), we obtain the required result.

Let  $P$  be a probability distribution defined by

$$P = \sum_{i=1}^{n_k} p_i \delta(x_i^k) \quad p_i = (h_{i-1} - h_i) x_i^k \quad p_{n_k} = h_{n_k} x_{n_k}^k = h_{n_k} L,$$

thus  $f_P = h$  and given the previous result, lemma 2.4 is proved.  $\square$

Given Lemma 2.4, we now prove Lemma 2.1.

*Proof of Lemma 2.1.* We first consider the case where  $\theta^{t_1} \lesssim \alpha(\theta) \lesssim \theta^{t_2}$  for small  $\theta$ . For  $\epsilon_n$  as in Theorem 2.1, define  $\theta_n$  as

$$\theta_n = \inf\{x, 1 - F_0(x) < \frac{\epsilon_n}{2n}\}.$$

Note that  $F_0$  is càdlàg, thus

$$F_0(\theta_n) \geq 1 - \epsilon_n/(2n) \text{ and } \forall y < \theta_n, 1 - F_0(y) > \epsilon_n/(2n). \quad (2.27)$$

. Using lemma 2.4 with  $L = \theta_n$ , we obtain that there exists a distribution  $P = \sum_{i=1}^{n_k} \delta_{x_i} p_i$  such that

$$KL(f_{0,n}, f_P) \leq \epsilon_n^2, \text{ and } \int f_{0,n} \log \left( \frac{f_{0,n}}{f_P} \right)^2 \lesssim \epsilon_n^2.$$

Note that  $f_P$  has support  $[0, \theta_n]$  and is such that  $f_P(\theta_n) > 0$ . Now, set  $m = n_k$  and consider  $P'$  the mixing distribution associated with  $\{m, x'_1, \dots, x'_m, p'_1, \dots, p'_m\}$  with  $\sum_{i=1}^m p'_i = 1$ . Define for  $1 \leq i \leq m-1$  the set  $U_i = [0 \vee (x_i - \epsilon_n^3/M, x_i + \epsilon_n^3/M]$  and  $U_m = (\theta_n, \theta_n + \epsilon_n(L - \theta_n) \wedge \epsilon_n^3/M]$ . Construct  $P'$  such that  $x'_i \in U_i$  and  $|P'(U_i) - p_i| \leq \epsilon^2 m^{-1}$ . We get

$$\forall t \in [0, \theta_n] \quad f'_P(t) > \frac{p'_m}{x'_m}.$$

Given that  $x'_m \in U_m$ , we get  $x'_m \leq \theta_n + \epsilon_n(L - \theta_n) \wedge \epsilon_n^3/M \lesssim \theta_n$  for  $n$  large enough. Note also that  $p'_m \geq p_m - \epsilon_n^2 m^{-1}$ . Given the construction of Lemma 2.4, we deduce

$$p_m \geq \frac{f_0(x_{i-1})}{1 + \mathcal{O}(\epsilon_n)} \gtrsim f_0(x_{i-1}),$$

for  $n$  large enough. Furthermore, given (2.27)

$$\forall z < \theta_n, \quad f_0(z)(L - z) \geq \int_z^L f_0(t) dt \geq \frac{\epsilon_n}{2n},$$

thus

$$\forall t \in [0, \theta_n] \quad f'_P(t) \gtrsim \frac{\frac{\epsilon_n}{2n} - \epsilon_n^2 m^{-1}}{\theta_n} \gtrsim \frac{\epsilon_n}{n},$$

and deduce that  $\|f_0/f_{P'}\|_\infty \lesssim \frac{n}{\epsilon_n}$  Lemma 8 from Ghosal and van der Vaart (2007) gives us that

$$\begin{aligned} \int_0^{\theta_n} f_0(x) \log \left( \frac{f_0}{f_{P'}} \right) (x) dx &\lesssim (\epsilon_n^2 + H^2(f_P, f_{P'})) (1 + |\log(\epsilon_n/n)|) \\ &\lesssim (\epsilon_n^2 + |f_P - f_{P'}|_1) (1 + |\log(\epsilon_n/n)|). \end{aligned}$$

Given the mixture representation (2.1) of  $f_0$  and  $f_P$ , we get

$$\begin{aligned} &(\epsilon_n^2 + |f_P - f_{P'}|_1) (1 + \log(n)) \\ &\lesssim \left( \epsilon_n^2 + \int_0^{\theta_n} \left| \sum \left( \frac{p_i}{x_i} - \frac{p'_i}{x'_i} \right) \mathbb{I}_{x \leq x_i} + \sum \frac{p_i}{x_i} (\mathbb{I}_{x \leq x_i} - \mathbb{I}_{x \leq x'_i}) \right| dx \right) (1 + \log(n)) \\ &\lesssim \left( \epsilon_n^2 + \sum \left| \frac{x_i}{x'_i} - 1 \right| p'_i + \sum |p'_i - p_i| + \sum \frac{p_i}{x_i} |x'_i - x_i| \right) (1 + |\log(n)|) \\ &\lesssim \epsilon_n^2 (1 + |\log(n)|). \end{aligned}$$

Generally speaking, denoting  $U_0 = [0, 1] \cap (\cup_{i=1}^m U_i)^c$  and  $\mathcal{N} = \{P', |P'(U_i) - p_i| \leq \epsilon_n^2 m^{-1}\}$  we obtain that for all  $P' \in \mathcal{N}$

$$\int_0^{\theta_n} f_0(x) \log \left( \frac{f_0}{f_{P'}} \right) (x) dx \lesssim \epsilon_n^2 (1 + |\log(n)|),$$

and similarly

$$\int_0^{\theta_n} f_0(x) \log \left( \frac{f_0}{f_{P'}} \right)^2 (x) dx \lesssim \epsilon_n^2 (1 + |\log(n)|)^2,$$

for  $\epsilon_n$  small enough. Note also that for all  $P' \in \mathcal{N}$  and  $n$  large enough, as before we get

$$\int_{\theta_n}^L f_{P'}(x) dx \lesssim \frac{\epsilon_n}{n}.$$

We now derive a control on  $k$ , the number of steps until  $\varepsilon_k \leq \epsilon_n^{3/2}$  in the construction of Lemma 2.4. At step  $k-1$ , we have  $\varepsilon_{k-1} \geq \epsilon_n^{3/2}$ . It is clear that for all  $j$ ,  $\varepsilon_j \leq 2^{-1/2} \varepsilon_{j-1}$ , thus

$$\begin{aligned} M^{1/2} L^{1/2} 2^{-(k-1)/2} &\geq \varepsilon_{k-1} \geq \epsilon_n^{3/2} \\ \log(M^{1/2} L^{1/2}) - (k-1) \frac{\log(2)}{2} &\geq \frac{3}{2} \log(\epsilon_n). \end{aligned}$$

Finally, we have

$$k \leq \frac{2}{\log(2)}(\log(M^{1/2}L^{1/2}) - \frac{3}{2}\log(\epsilon_n)) + 1. \quad (2.28)$$

We can then get a lower bound for  $\Pi[\mathcal{N}]$  and, given that for  $\epsilon_n$  small enough and  $n$  large enough, we have

$$\mathcal{N} \subset S_n(\epsilon_n, \theta_n),$$

we can deduce a lower bound for  $\Pi(S_n(\epsilon_n, \theta_n))$ . For the Type 1 prior, we have similarly to Ghosal et al. (2000)

$$\begin{aligned} \Pi[\mathcal{N}] &= \Pr(\mathcal{D}(A\alpha(U_0), \dots, A\alpha(U_{n_k})) \in [p_i \pm \epsilon_n^2/n_k]) \\ &\geq \frac{\Gamma(A)}{\prod_i \Gamma(A\alpha(U_i))} \prod_j \int_{(p_i - \epsilon_n^2/n_k) \wedge 0}^{(p_i + \epsilon_n^2/n_k)} x_j^{A\alpha(U_j)-1} dx_j. \end{aligned}$$

Given condition C1, we have

$$\alpha(U_i) \geq \int_{U_i} \alpha_0 \theta^{t_1} d\theta,$$

thus

$$\alpha(U_i) \geq 2\epsilon_n^3 \alpha_0 x_i^{t_1}.$$

for  $n$  large enough and  $\epsilon$  sufficiently small we have as in Lemma 6.1 of Ghosal et al. (2000)

$$\Pi(\mathcal{N}) \gtrsim \exp\{C_1 n_k \log(\epsilon)\}.$$

Note that given (2.25),  $n_k \lesssim \epsilon_n^{-1}$  which gives the desired result. For the Type 2 prior, we write

$$\mathcal{N}' = \left\{ P' = \sum_{j=1}^{n_k} p'_j \delta_{x'_j}, |p'_j - p_j| \leq \epsilon^2/n_k, |x'_j - x_j| \leq \epsilon_n^3 \right\} \subset S_n(\epsilon_n, \theta_n),$$

we then deduce a lower bound for  $\Pi[S_n(\epsilon_n, \theta_n)]$

$$\begin{aligned} \Pi[\mathcal{N}'] &\geq Q(K = n_k) \prod_{j=1}^{n_k} n_k^{-n_k} c^{n_k} \int_{\max(0, p_i - \epsilon^2/n_k)}^{p_i + \epsilon^2/n_k} w_j^{a_j} dw_j \prod_{j=1}^{n_k} \alpha(U_i) \\ &\geq \exp\left\{-cn_k \log n_k + \sum \log(\alpha(U_i)) + n_k \log(c) - n_k \log(n_k) + \sum a_j \log(2\epsilon^2/n_k)\right\} \\ &\gtrsim \exp\{C'_1 \epsilon^{-1} \log(\epsilon)\}. \end{aligned}$$

We now consider the case where  $e^{-a_1/\theta} \leq \alpha(\theta) \leq e^{-a_2/\theta}$  if  $\theta$  is close to 0 and  $\sup_{x \in [0, \delta]} |f'_0(x)| \leq C_0$ . We have that for  $n$  large enough and  $C > 0$ , a constant depending on  $f_0$ ,  $f_0(0) - f_0(\epsilon_n) \leq C\epsilon_n$ . Following Lemma 2.4, we can construct a piecewise constant approximation of  $f_0$  on  $[\delta, L]$ . On  $[0, \delta]$ , consider the regular partition with  $\lfloor \epsilon_n^{-1} \rfloor$  points and the piecewise constant approximation of  $f_0$  defined as before (i.e.  $f_i = f_0(x_{i-1})$ ). Again, this approximation can be identified with a measure  $P$ . Given the assumptions on  $f_0$  we immediately get that  $KL(f_0, f_P) \lesssim \epsilon_n^2$ .

Consider the same sets  $\mathcal{N}$  as before, with the same partitions  $U_1, \dots, U_n$ . Using similar computations as in Lemma 6.1 of Ghosal et al. (2000) we get that

$$\Pi(\mathcal{N}) \geq \exp \left\{ C_1(n_k + \epsilon_n^{-1}) \log(\epsilon_n) + \sum \log(\alpha(U_i)) \right\}$$

For the  $U_i$  included in  $[\delta, L]$  we have  $\alpha(U_i) \gtrsim \epsilon_n^{3/2}$ . For the  $U_i$  included in  $[0, \delta]$  we have  $\alpha(U_i) \gtrsim \epsilon_n \exp \{-a/(i\epsilon_n)\}$ , which gives

$$\sum \alpha(U_i) \lesssim -\epsilon_n^{-1} \log(n)$$

We end the proof using similar argument as before. □

### 2.4.2 Proof of Lemma 2.2

The proof of Lemma 2.2 is straightforward and comes directly from C1 and C2.

*Proof.* Recall that given (2.1),  $f(0) = \int_{[0,1]} \frac{1}{\theta} dP(\theta)$ . Then

$$\Pi \left[ \int_0^1 \frac{1}{\theta} dP(\theta) \geq M_n \right] = \Pi \left[ \int_0^{2M_n^{-1}} \frac{1}{\theta} dP(\theta) + \int_{2M_n^{-1}}^1 \frac{1}{\theta} dP(\theta) \geq M_n \right].$$

Note that

$$\int_{2M_n^{-1}}^1 \frac{1}{\theta} dP(\theta) \leq M_n/2 \int_{2M_n^{-1}}^1 dP(\theta) \leq M_n/2.$$

Thus the set  $\{P, \int_0^{2M_n^{-1}} \theta^{-1} dP(\theta) \geq M_n/2\}$  contains  $\mathcal{F}_n^c$  and

$$\begin{aligned} \Pi[\mathcal{F}_n^c] &\leq \Pi \left[ \int_0^{2M_n^{-1}} \frac{1}{\theta} dP(\theta) > M_n/2 \right] \\ &\leq 2M_n^{-1} E \left[ \int_0^{2M_n^{-1}} \frac{1}{\theta} dP(\theta) \right], \end{aligned}$$

using Markov inequality. Then for a Type 1 prior when  $n$  large enough

$$\begin{aligned}\Pi[\mathcal{F}_n^c] &\leq 2M_n^{-1} \int_0^{2M_n^{-1}} \frac{1}{\theta} \alpha(\theta) d\theta \\ &\leq 2M_n^{-1} \int_0^{2M_n^{-1}} \theta^{t_2-1} d\theta = \frac{(2M_n^{-1})^{t_2+1}}{t_2} = Ce^{-cn^{1/3} \log(n)^{2/3}}.\end{aligned}$$

For a Type 2 prior, we have that

$$\begin{aligned}\Pi[\mathcal{F}_n^c] &\leq \sum_{h=1}^{\infty} Q(K=h) \pi_h \left[ \min_{j \leq h} x_j \leq M_n^{-1} \right] \\ &\leq \left( \sum_{h=1}^{\infty} h Q(K=h) \right) \alpha([0, M_n^{-1}]) \\ &\leq C' e^{-cn^{1/3} \log(n)^{2/3}}.\end{aligned}$$

□

## 2.5 Adaptation of Theorem 4 of Rivoirard et al. (2012)

This Theorem is a slight modification of Theorem 2.9 of Ghosal et al. (2000). The main difference lies in the handling of the denominator  $D_n$  in

$$\Pi(f : d(f_0, f) \geq J_{0,n} \epsilon_n | \mathbf{X}^n) = \frac{\int_{d(f, f_0) \geq J_{0,n} \epsilon_n} \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} d\Pi(f)}{\int \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} d\pi(f)} = \frac{N_n}{D_n},$$

as in general, it requires a lower bound on the prior mass of Kullback Leibler neighborhood of  $f_0$ . Here we prove that under condition (2.16) we have for some constants  $c, C > 0$

$$P_0^n(D_n < ce^{-Cn\epsilon_n^2}) = o(1).$$

Let  $l_n(f)$  be the log likelihood associated with  $f$  and define  $\Omega_n = \{(f, \mathbf{X}^n), l_n(f) - l_n(f_0) > -C_1 n \epsilon_n^2\}$  for some constant  $C_1 > 0$ . Define also  $A_n = \{\mathbf{X}^n, \forall i X_i \leq \theta_n\}$ . We thus have

$$D_n \geq e^{-C_1 n \epsilon_n^2} \int_{S_n(\epsilon_n, \theta_n)} \mathbb{I}_{\Omega_n} d\Pi(f) = e^{-C_1 n \epsilon_n^2} \Pi(S_n(\epsilon_n, \theta_n) \cap \Omega_n).$$

Note that given (2.16) we have that there exists  $\rho > 0$  such that for  $n$  large enough  $e^{-C_2 n \epsilon_n^2} \Pi(S_n(\epsilon_n, \theta_n) > \rho)$ . We now write

$$\begin{aligned} P_0^n(D_n < e^{-C n \epsilon_n^2}) &\leq P_0^n\left(e^{(C-C_1)n\epsilon_n^2} \Pi(S_n(\epsilon_n, \theta_n) \cap \Omega_n) < c\right) \\ &\leq P_0^n\left(e^{(C-C_1-C_2)n\epsilon_n^2} \Pi(S_n(\epsilon_n, \theta_n) \cap \Omega_n) < \frac{c}{\rho} \Pi(S_n(\epsilon_n, \theta_n))\right) \\ &\leq P_0^n\left(\Pi(S_n(\epsilon_n, \theta_n) \cap \Omega_n^c) > \left(1 - e^{-(C-C_1-C_2)n\epsilon_n^2} \frac{c}{\rho}\right) \Pi(S_n(\epsilon_n, \theta_n))\right) \\ &\leq \frac{2 \int_{S_n(\epsilon_n, \theta_n)} P_0^n(\Omega_n^c) d\Pi(f)}{\Pi(S_n(\epsilon_n, \theta_n))}. \end{aligned}$$

For all  $f \in S_n(\epsilon_n, \theta_n)$  we compute

$$\begin{aligned} m_n &= E_0^n(l_n(f_0) - l_n(f) \mathbb{I}_{A_n}) \\ &= n F_0(\theta_n)^{n-1} \int_0^{\theta_n} f_0 \log\left(\frac{f_0(x)}{f(x)}\right) dx \\ &= n F_0(\theta_n)^n \left( KL(f_{0,n}, f_n) + \log\left(\frac{F_0(\theta_n)}{F(\theta_n)}\right) \right) \\ &\leq C_3 n \epsilon_n^2, \end{aligned}$$

and

$$\begin{aligned} P_0^n(\Omega_n^c) &= P_0^n(l_n(f) - l_n(f_0) < -C_1 n \epsilon_n^2) \\ &= P_0^n(\{l_n(f) - l_n(f_0) < -C_1 n \epsilon_n^2\} \cap A_n) + o(1) \\ &\leq P_0^n(\{l_n(f_0) - l_n(f) - m_n > (C_1 - C_3) n \epsilon_n^2\} \cap A_n) + o(1) \\ &\leq \frac{E_0^n(\{l_n(f_0) - l_n(f) - m_n\} \mathbb{I}_{A_n})^2}{(C_1 - C_3)^2 (n \epsilon_n^2)^2} + o(1). \end{aligned}$$

We then compute for  $C_5$  and  $C_6$  some fixed constants

$$\begin{aligned} v_n &= E_0^n(\{l_n(f_0) - l_n(f) - m_n\} \mathbb{I}_{A_n})^2 \\ &= (F_0(\theta_n))^{n-1} \left( n \int_0^{\theta_n} f_0 \log^2\left(\frac{f_0(x)}{f(x)}\right) dx + n(n-1) \left( \int_0^{\theta_n} f_{0,n} \log\left(\frac{f_0(x)}{f(x)}\right) dx \right)^2 - m_n^2 \right) \\ &= (F_0(\theta_n))^{n-1} \left( n \int_0^{\theta_n} f_0 \log^2\left(\frac{f_0(x)}{f(x)}\right) dx + \frac{n-1}{n} F_0(\theta_n)^{-2n+2} m_n^2 - m_n^2 \right) \\ &\leq n F_0(\theta_n)^n \int_0^{\theta_n} f_{0,n} \log^2\left(\frac{f_0(x)}{f(x)}\right) dx + \frac{n-1}{n} m_n^2 F_0(\theta_n)^{n-1} (F_0(\theta_n)^{-2n+2} - 1) \\ &\leq C_5 n \epsilon_n^2 + C_6 (n \epsilon_n^2)^2 \epsilon_n. \end{aligned}$$

We finally obtain that for all  $f \in S_n(\epsilon_n, \theta_n)$ ,  $P_0^n(\Omega_n^c) = o(1)$ . We end the proof using similar arguments as in Ghosal et al. (2000).

## 2.6 Discussion

In this chapter, we obtain an upper bound for the concentration rate of the posterior distribution under monotonicity constraints. This is of interest as in this model, the standard approach based on the seminal paper of Ghosal et al. (2000) cannot be applied directly. We prove that the concentration rate of the posterior is (up to a  $\log(n)$  factor) the minimax estimation rate  $(n/\log(n))^{-1/3}$  for standard losses such as  $L_1$  or Hellinger.

We also prove that the posterior distribution is consistent for the pointwise loss at any point of the support and for the sup norm loss. Studying asymptotic properties for these losses is difficult in general as the usual approach are well suited for losses that are related to the Hellinger metric. Obtaining more refined results on the asymptotic behaviour of the posterior distribution will require refined control of the likelihood which in the case of nonparametric mixture models is a difficult task.

## Bibliography

- Balabdaoui, F. and Wellner, J. A. (2007). Estimation of a  $k$ -monotone density: limit distribution theory and the spline connection. *Ann. Statist.*, 35(6):2536–2564.
- Brunner, L. J. and Lo, A. Y. (1989). Bayes methods for a symmetric unimodal density and its mode. *Ann. Statist.*, 17(4):1550–1566.
- Castillo, I. (2013). On bayesian supremum norm contraction rates. *arXiv preprint arXiv:1304.1761*.
- Donoho, D. L. and Liu, R. C. (1991). Geometrizing rates of convergence, ii. *The Annals of Statistics*, pages 633–667.
- Durot, C., Kulikov, V. N., Lopuhaä, H. P., et al. (2012). The limit distribution of the  $l_{\infty}$ -error of grenander-type estimators. *The Annals of Statistics*, 40(3):1578–1608.
- Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent advances in statistics*, pages 287–302. Academic Press, New York.
- Ghosal, S., Ghosh, J., and van der Vaart, A. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.
- Ghosal, S. and van der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35(2):697–723.
- Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5):1233–1263.

- Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.*, 38(2):1122–1170.
- Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.*, 39:125–153 (1957).
- Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 539–555, Belmont, CA. Wadsworth.
- Groeneboom, P. (1986). Some current developments in density estimation. In *Mathematics and computer science (Amsterdam, 1983)*, volume 1 of *CWI Monogr.*, pages 163–192. North-Holland, Amsterdam.
- Hoffmann, M., Rousseau, J., and Schmidt-Hieber, J. (2013). On adaptive posterior concentration rates. *arXiv preprint arXiv:1305.5270*.
- Khazaei, S., Rousseau, J., and Balabdaoui, F. (2010). Bayesian Nonparametric Inference of decreasing densities. In *42èmes Journées de Statistique*, Marseille, France France.
- Kruijer, W., Rousseau, J., and van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.*, 4:1225–1257.
- Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Stat.*, 12:351–357.
- Prakasa Rao, B. L. S. (1970). Estimation for distributions with monotone failure rate. *Ann. Math. Statist.*, 41:507–519.
- Reynaud-Bouret, P., Rivoirard, V., and Tuleau-Malot, C. (2011). Adaptive density estimation: a curse of support? *J. Statist. Plann. Inference*, 141(1):115–139.
- Rivoirard, V., Rousseau, J., et al. (2012). Bernstein–von mises theorem for linear functionals of the density. *The Annals of Statistics*, 40(3):1489–1523.
- Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester.
- Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.*, 38(1):146–180.
- Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(5):689–710.
- Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive bayesian multivariate density estimation with dirichlet mixtures. *Biometrika*, 100(3):623–640.
- Sun, J. and Woodroffe, M. (1996). Adaptive smoothing for a penalized NPMLE of a non-increasing density. *J. Statist. Plann. Inference*, 52(2):143–159.
- van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empir-*

- ical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.
- Williamson, R. E. (1956). Multiply monotone functions and their Laplace transforms. *Duke Math. J.*, 23:189–207.
- Wu, Y. and Ghosal, S. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electron. J. Stat.*, 2:298–331.



## Chapter 3

# Bayesian testing for monotonicity

“Every day is a new day. It is better to be lucky. But I would rather be exact. Then when luck comes you are ready.”

– **Ernest Hemingway**, *The old man and the sea*.

### Résumé

Nous proposons un test bayésien non paramétrique de décroissance d’une fonction dans le modèle de régression gaussien. Dans ce cadre, outre le fait que les deux hypothèses sont non-paramétriques, l’hypothèse nulle est incluse dans l’alternative. Il s’agit donc d’un cas de test particulièrement difficile. En outre dans ce cas, l’approche usuelle par le facteur de Bayes n’est pas consistante. Nous proposons donc une approche alternative reprenant les idées d’approximation d’une hypothèse ponctuelle par un intervalle. Nous prouvons que pour une large famille de lois a priori, le test proposé est consistant et sépare les hypothèses à la vitesse minimale. De plus notre procédure est facile à implémenter et à mettre en œuvre. Nous étudions ensuite son comportement sur des données simulées et comparons les résultats avec les méthodes classiques existantes dans la littérature. Pour chacun des cas considérés, nous obtenons des résultats au moins aussi bons que les méthodes existantes, et les surpassons pour un certain nombre de cas.

## 3.1 Introduction

### 3.1.1 Modelling with monotone constraints

Shape constraints models, and monotone constraints models in particular, are of growing interest in the nonparametric field. There is a wide literature on the problem of estimating monotone functions. Groeneboom (1985), Prakasa Rao (1970) and Robertson et al. (1988) among others study the nonparametric maximum likelihood estimator of monotone densities, Lo (1984), Brunner and Lo (1989), and Salomond (2013) study some posterior distribution in a Bayesian approach. Barlow et al. (1972) and Mukerjee (1988) proposed a shape constraint estimators of monotonic regression functions. These methods are widely applied in practice. Bornkamp and Ickstadt (2009) consider monotone function when modeling the response to a drug as a function of the dose and Neittaanmäki et al. (2008) use a monotone representation for environmental data.

In this chapter we propose a procedure to test for monotonicity constraints in the Gaussian regression model

$$Y_i = f(i/n) + \sigma\epsilon_i, \epsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, 1), \sigma > 0, i = 1, \dots, n, \quad (3.1)$$

and, with  $\mathcal{F}(K)$  being the set of all monotone functions uniformly bounded by  $K$ , we test

$$H_0 : f \in \mathcal{F}(K), \text{ versus } H_1 : f \notin \mathcal{F}(K). \quad (3.2)$$

Here both the null and the alternative are nonparametric hypotheses. The problem of testing for monotonicity has already been addressed in the frequentist literature and a variety of approaches have been considered. Baraud et al. (2005) use projections of the regression function on the sets of piecewise constant function on a collection of partition of support of  $f$ . Their test rejects monotonicity if there is at least one partition such that the estimated projection is too far from the set of monotone functions. Another approach, considered in Hall and Heckman (2000) and Ghosal et al. (2000) among others, is to test for negativity of the derivative of the regression function. However this requires some assumptions on the regularity of the regression function under the null hypothesis that could be avoided. In a recent paper Akakpo et al. (2014) propose a procedure that detects local departure from monotonicity, and study very precisely its asymptotic properties.

Here, we propose a Bayesian approach to this problem, which to the author's knowledge has only receive little consideration. Scott et al. (2013) consider a Bayesian test for monotonicity based on constrained spline. Their approach require smoothness assumptions on the regression function under the alternative, which we avoid here. We only consider the case where  $\mathcal{F}(K)$  is the set of monotone non increasing functions uniformly bounded by  $K$ , but a similar approach could

be used when considering the set of monotone increasing. The most common approach to testing in a Bayesian setting is the Bayes Factor. Here however, we see that this method has drawbacks and seems to have poor performances, hence we propose a modification of the Bayes factor.

### 3.1.2 The Bayes factor approach

The standard Bayesian answer to the testing problem (3.2) related with the 0 – 1 loss is the Bayes factor

$$B_{0,1} = \frac{\Pi \{f \in \mathcal{F}(K) \mid Y^n\}}{\Pi \{f \notin \mathcal{F}(K) \mid Y^n\}} \frac{1 - \Pi \{\mathcal{F}(K)\}}{\Pi \{\mathcal{F}(K)\}}.$$

This approach to Bayesian testing is easy to understand as posterior probability of the considered hypotheses have a simple interpretation.

In this chapter we consider a prior on piecewise constant functions.

$$f = \sum_{i=1}^k \mathbb{I}_{[(i-1)/k, i/k)} \omega_i, \quad d\Pi(f) = \pi(k) \pi(\omega_1, \dots, \omega_k \mid k) d\lambda_k(\omega_1, \dots, \omega_k) d\nu(k),$$

where  $\lambda_k$  is the Lebesgue measure on  $\mathbb{R}^k$  and  $\nu$  the counting measure on  $\mathbb{N}$ . These prior are common in the Bayesian nonparametric literature. Furthermore for the problem of estimating monotone non increasing densities, related priors have been proved to lead to the minimax concentration rate over  $\mathcal{F}(K)$  in Salomond (2013).

In our case, the Bayes factor seems to give poor results in practice. The reason behind this is that when  $f$  has flat parts, it becomes difficult to detect monotonicity due to estimation uncertainty. For instance when considering the function  $f = 0$  the Bayes Factor does not seem to give a credible answer. As an illustration, Figure 3.1 gives the histogram constructed from 100 draws of data with  $f = 0$  and  $n = 100$ . Bayes Factor smaller than 0 indicates that the function is not monotone non increasing. It appears that for these runs, the Bayes Factor is rather small and that for a non negligible proportion of samples the log Bayes Factor is negative. Thus the answers given by the Bayes Factor are not satisfying in this case.

### 3.1.3 An alternative approach

To tackle this issue of constructing a test robust to flat parts, we change the formulation of our test into

$$H_0^a : \tilde{d}\{f, \mathcal{F}(K)\} \leq \tau \quad \text{versus} \quad H_1^a : \tilde{d}\{f, \mathcal{F}(K)\} > \tau \quad (3.3)$$

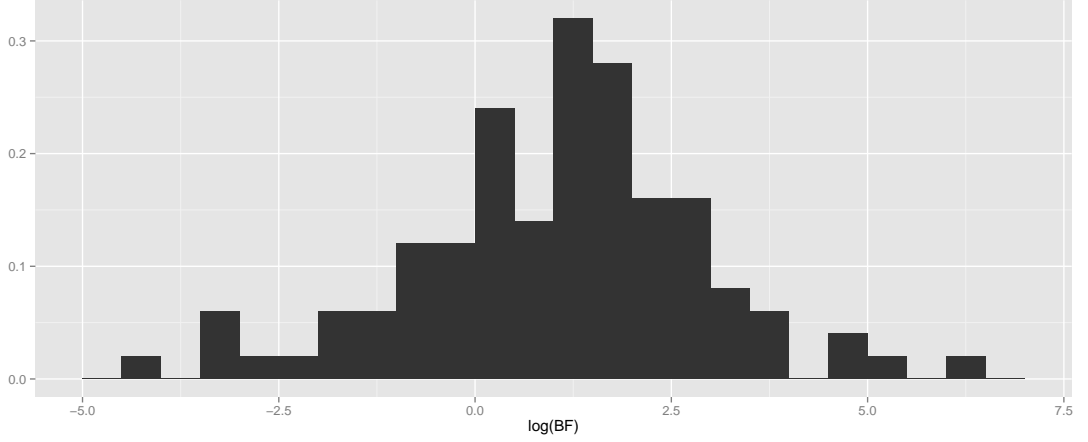


Figure 3.1: 100 simulation of the log Bayes Factor  $B_{0,1}$  for  $f = 0$  and  $n = 100$

where  $\tilde{d}(f, \mathcal{F}(K)) = \inf_{g \in \mathcal{F}(K)} \tilde{d}(f, g)$  and  $\tilde{d}$  is a metric or a semi-metric and  $\tau$  is a threshold. This idea is similar to the one proposed in Rousseau (2007) for the approximation of a point null hypothesis by an interval hypothesis testing. Here again we consider the 0 – 1 loss with weight  $\gamma_0, \gamma_1$  so that the Bayesian decision is given by

$$\delta_n^\pi = \begin{cases} 0 & \text{if } \Pi \left[ \tilde{d}\{f, \mathcal{F}(K)\} \leq \tau | Y_n \right] \geq \frac{\gamma_0}{\gamma_0 + \gamma_1} \\ 1 & \text{otherwise} \end{cases} \quad (3.4)$$

The threshold  $\tau$  can be calibrated a priori by a prior knowledge on the tolerance to approximate monotonicity. In practice such an a priori calibration is not always feasible. We therefore propose in this chapter an automatic calibration of  $\tau$ . In absence of prior information on the threshold, it is natural to have  $\tau$  depending on  $n$ , since the more data, the more precise we can afford to be. A least requirement will be that the test described in (3.3) is asymptotically equivalent to the test (3.2). Hence a calibration of  $\tau$  such that our test is consistent, that is for all  $\rho > 0$  and  $d(\cdot, \cdot)$  a metric or a semi-metric, potentially different from  $\tilde{d}$ ,

$$\begin{aligned} \sup_{f \in \mathcal{F}(K)} \mathbb{E}_f^n(\delta_n^\pi) &= o(1) \\ \sup_{f, d\{f, \mathcal{F}(K)\} > \rho} \mathbb{E}_f^n(1 - \delta_n^\pi) &= o(1). \end{aligned} \quad (3.5)$$

To understand the effectiveness of the threshold induced by our approach, we study the minimum separation rate of our test which is the minimum value  $\rho = \rho_n$  such that (3.5) is still valid. Small  $\rho_n$  implies that the test is able to detect very

small departure from the null. We thus want our calibrated threshold to induce the smallest separation rate.

Form a practical point on view, this procedure will be easy to implement as it will only require sampling under the posterior distribution which is made easy by our choice of prior. This is a great advantage compared to the frequentist tests proposed in the literature as they require in general heavy computations.

We thus propose a procedure which although being a Bayesian answer to the problem (3.3), is also asymptotically an answer to the problem (3.2). Moreover, our procedure is automatic and easy to implement. The construction of the test is presented in section 3.2 and its asymptotic properties are discussed in Section 3.2.2. In Section 3.2.3 we propose a way to calibrate the hyperparameters of the prior rendering the procedure fully automatic. We then run our test on simulated data in section 3.3 and on real environmental data in section 3.4. A general discussion is provided in section 3.7.

## 3.2 Construction of the test

### 3.2.1 The testing procedure

We first propose a choice for  $\tilde{d}\{f, \mathcal{F}(K)\}$  which measures the distance between the regression function  $f$  and the set  $\mathcal{F}(K)$  and a way to calibrate the threshold  $\tau$  in situation where prior information is not available. This is done such that by answering the problem (3.3) we give a good answer to the problem (3.2). We then propose a specific family of prior that will speed up the computations together with a choice for the hyperparameters based on heuristics.

As presented in section 3.1.1, monotone non increasing functions are well approximated by stepwise constant functions. Let  $\mathcal{G}_k$  be the set of piecewise constant functions with  $k$  pieces on the partition  $\{[0, 1/k), \dots, [(k-1)/k, 1]\}$  so that each function in  $\mathcal{G}_k$  will be written

$$f_{\omega,k}(\cdot) = \sum_{i=1}^k \omega_i \mathbb{I}_{[(i-1)/k, i/k)}(\cdot), \quad \omega = (\omega_1, \dots, \omega_k) \in \mathbb{R}^k. \quad (3.6)$$

We assume that the data  $Y^n = (Y_1, \dots, Y_n)$  is generated by model (3.1), where the residual variance  $\sigma^2$  is unknown. We then build a prior on  $(f, \sigma)$  taking a prior on  $k$  and building a prior on each submodels  $\mathcal{G}_k$ . We define

$$\Pi(\omega, \sigma, k) := \Pi(k)\Pi(\sigma|k)\Pi(\omega|\sigma, k).$$

First with this choice of prior we have generally speaking  $\pi(\mathcal{F}(K)) > 0$ . Furthermore, if the true regression function  $f_0$  is in  $\mathcal{F}(K)$  then the piecewise constant

function in  $\mathcal{G}_k$  of the form (3.6) which minimizes the Kullback Leibler divergence with  $f_0$  will also be in  $\mathcal{F}(K)$  for all  $k$ . We consider the following discrepancy measure  $\tilde{d}(\cdot, \cdot)$  in (3.3) between  $f_{\omega,k} \in \mathcal{G}_k$  and  $\mathcal{F}(K)$ ,

$$\tilde{d}\{f_{\omega,k}, \mathcal{F}(K)\} = H(\omega, k) = \max_{k \geq j \geq i \geq 1} (\omega_j - \omega_i). \quad (3.7)$$

From (3.7) it appears that  $f_{\omega,k}$  is in  $\mathcal{F}(K)$  if and only if  $\tilde{d}\{f_{\omega,k}, \mathcal{F}(K)\} = 0$ . Here the discrepancy  $\tilde{d}$  corresponds to the sup norm between  $f_{\omega,k}$  and the set of monotone non increasing functions. The idea of the calibration is the following. In the model  $\mathcal{G}_k$ , the a posteriori uncertainty for estimating  $\omega = (\omega_1, \dots, \omega_k)$  is of order  $\sqrt{k/n}$ . Hence any monotone non increasing function  $f_{\omega,k}$  such that for all  $j > i$ ,  $\omega_i \geq \omega_j - O(\sqrt{k/n})$  might be detected as possibly monotone non increasing. We thus choose a threshold  $\tau_n^k$  for each model  $\mathcal{G}_k$ . We then compare  $H(\omega, k)$  with some positive threshold depending on  $n$  and  $k$  and then calibrate  $\tau_n^k$  such that our procedure is consistent. To evaluate the effectiveness of the threshold, we consider Hölderian alternatives, following what is done in the frequentist literature,

$$f \in \mathcal{H}(\alpha, L) = \{f, [0, 1] \rightarrow \mathbb{R}, \forall x, y \in [0, 1]^2 |f(y) - f(x)| \leq L|y - x|^\alpha\},$$

for some constant  $L > 0$  and a regularity parameter  $\alpha \in (0, 1]$ . We study the separation rate of our procedure and compare it with the minimax separation rate  $n^{-\alpha/(2\alpha+1)}$ .

### 3.2.2 Theoretical results

The following Theorem provides a way to calibrate  $\tau_n^k$ . It also gives an upper bound for the minimal separation rate with respect to the distance  $d_\infty(\cdot, \cdot)$  defined as

$$d_\infty(f, g) = \sup_{x \in [0, 1]} \{|f(x) - g(x)|\}$$

Consider prior of the form

$$\frac{d\Pi_\omega}{d\lambda_k} = g^{\otimes k}, \quad \frac{d\Pi_\sigma}{d\lambda_1} = \pi_\sigma, \quad \frac{d\Pi_k}{d\nu} = \pi_k,$$

where  $\lambda_k$  is the Lebesgue measure on  $\mathbb{R}^k$ , which satisfies the following conditions :

- C1** the density  $\pi_\sigma$  is continuous and  $\pi_\sigma(\sigma) > 0$  for all  $\sigma \in (0, \infty)$ ,
- C2** the density  $g$  is continuous and puts mass on all  $\mathbb{R}$ . Furthermore,  $g$  is such that there exists a constant  $c_g$  such that for all  $K > 0$ , for all  $z > 0$ , for all  $l \in \mathbb{N}$ , for all sequence  $u$  that goes to 0,

$$\sup_{|x_0| \leq K} \frac{\int \mathbb{I}[lzu \leq |x - x_0| \leq (l+1)zu] g(x) dx}{\sup_{|x_0| \leq K} \int \mathbb{I}[|x - x_0| \leq zu] g(x) dx} \leq u^{-c_g},$$

**C3**  $\pi_k$  is such that there exists positive constants  $C_d$  and  $C_u$  such that

$$e^{-C_d k L(k)} \leq \pi_k(k) \leq e^{-C_u k L(k)} \quad (3.8)$$

where  $L(k)$  is either  $\log(k)$  or 1.

The condition **C1** and **C2** are mild and are satisfied for a large variety of distributions. In section 3.2.3 we will take  $g$  to be a Gaussian density and  $\pi_\sigma$  to be an inverse gamma. Simple algebra shows that for this choice of prior, both conditions are satisfied. **C3** is a usual condition when considering mixture models with random number of components, see e.g. Rousseau (2010) and is satisfied by Poisson or Geometric distribution for instance. We then have the following control on our test:

**Theorem 3.1.** *Under the assumptions **C1** to **C3**, for a fixed constant  $M_0 > 0$ , setting  $\tau = \tau_n^k = M_0 \{k \log(n) n^{-1}\}^{1/2}$  and  $\delta_n^\pi$  the testing procedure defined in (3.4), for all  $K > 0$  then there exists some  $M > 0$  such that for all  $\alpha \in (0, 1]$*

$$\begin{aligned} \sup_{f \in \mathcal{F}(K)} E_f^n(\delta_n^\pi) &= o(1) \\ \sup_{f, d_\infty\{f, \mathcal{F}(K)\} > \rho, f \in \mathcal{H}(\alpha, L)} E_f^n(1 - \delta_n^\pi) &= o(1) \end{aligned} \quad (3.9)$$

for all  $\rho > \rho_n(\alpha) = M \{n / \log(n)\}^{-\alpha/(2\alpha+1)} v_n$  where  $v_n = 1$  when  $L(k) = \log(k)$  and  $v_n = \{\log(n)\}^{1/2}$  when  $L(k) = 1$ .

Neither the prior nor the hyperparameters depends on the regularity  $\alpha$  of the regression function under the alternative. Moreover for all  $\alpha \in (0, 1]$ , the separation rate  $\rho_n(\alpha)$  is the minimax separation rate up to a  $\log(n)$  term. Thus our test is almost minimax adaptive. The  $\log(n)$  term seems to follow from our definition of the consistency where we do not fix a level for the Type I or Type II error contrariwise to the frequentist procedures. The conditions on the prior are quite loose, and are satisfied in a wide variety of cases. The constant  $M_0$  does not influence the asymptotic behaviour of our test but has a great influence in practice for finite  $n$ . A way of choosing  $M_0$  is given in section 3.2.3.

The proof of Theorem 4.1 is given in Section 3.5, we sketch here the main ideas. We approximate the true regression function  $f_0$  in each submodel  $\mathcal{G}_k$  by  $f_{\omega^0, k}$  that minimizes the Kullback-Leibler divergence with  $f_0$ . We have a close form expression for  $\omega^0 = (\omega_1^0, \dots, \omega_k^0)$  given by

$$\omega_i^0 = n_i^{-1} \sum_{j, j/n \in [(i-1)/k, i/k)} f_0(j/n), \quad n_i = \text{Card} \{j, j/n \in [(i-1)/k, i/k)\} \quad (3.10)$$

thus  $f_{\omega^0, k}$  belongs to  $\mathcal{F}$  for all  $k$  when  $f_0 \in \mathcal{F}$ . To prove the first part of (3.9), we bound  $H(\omega, k) \leq 2 \max |\omega_i - \omega_i^0|$  if  $f_0 \in \mathcal{F}$  so that the threshold  $\tau_n^k$  needs to be

as large as the posterior concentration rate of  $\omega$  to  $\omega^0$  in the misspecified model  $\mathcal{G}_k$ . Then to prove the second part of (3.9) when  $\rho = \rho_n(\alpha)$ , we bound from below  $H(\omega, k)$  by  $H(\omega^0, k) - 2 \max |\omega_i - \omega_i^0|$  which implies a constraint on the separation rate of the test to ensure that uniformly over  $d_n(f_0, \mathcal{F}) \geq \rho_n(\alpha)$  and  $f \in \mathcal{H}(\alpha, L)$  we have  $H(\omega, k) > \tau_n^k$ .

### 3.2.3 A choice for the prior in the non informative case

Conditions on the prior in Theorem 4.1 are satisfied for a wide variety of distributions. However, when no further information is available, some specific choices can ease the computations and lead to good results in practice. We present in this section such a specific choice for the prior and a way to calibrate the hyperparameters. We also fix  $\gamma_0 = \gamma_1 = 1/2$  in the definition of  $\delta_n^\pi$ .

A practical default choice is the usual conjugate prior, given  $k$ , i.e. a Gaussian prior on  $\omega$  with variance proportional to  $\sigma^2$  and an Inverse Gamma prior on  $\sigma^2$ . This will considerably accelerate the computations as sampling under the posterior is then straightforward. Condition (3.8) on  $\pi_k$  is satisfied by the two classical distributions on the number of parameters in a mixture model, namely the Poisson distribution and the Geometric distribution. It seems that choosing a Geometric distribution is more appropriate as it is less spiked. We thus choose

$$\Pi = \begin{cases} k \sim \text{Geom}(\lambda) \\ \sigma^2 | k \sim IG(a, b) \\ \omega_i | k, \sigma \stackrel{iid}{\sim} \mathcal{N}(m, \sigma^2/\mu) \end{cases} \quad (3.11)$$

Standard algebra leads to a close form for the posterior distribution up to a normalizing constant. Recall that  $n_j = \text{Card} \{i, i/n \in [(j-1)/k, j/k]\}$ , we denote

$$\tilde{b}_k = b + \frac{1}{2} \sum_{j=1}^k \left\{ \sum_{i, i/n \in I_j} (Y_i - \bar{Y}_j)^2 + \frac{n_j \mu}{n_j + \mu} (\bar{Y}_j - m)^2 \right\},$$

where  $\bar{Y}_j$  is the empirical mean of the  $Y_l$  on the set  $\{l, l/n \in [(j-1)/n, j/n]\}$ , we have

$$\pi_k(k | Y^n) \propto \pi(k) \tilde{b}_k^{-(\alpha + n/2)} \mu^{k/2} \prod_{j=1}^k (n_j + \mu)^{-1/2}$$

We can thus compute the posterior distribution of  $k$  up to a constant. To sample from  $\pi_k$  we use a random walk Hasting-Metropolis algorithm, see Robert and Casella

(2004). We then compute the posterior distribution of  $\omega$  and  $\sigma$  given  $k$

$$\begin{aligned}\sigma^2|k, Y^n &\sim IG(a + n/2, \tilde{b}_k) \\ \omega_j|k, \sigma^2, Y^n &\stackrel{ind.}{\sim} \mathcal{N}\left(\frac{m\mu + n_j\bar{Y}_j}{n_j + \mu}, \frac{\sigma^2}{n_j + \mu}\right).\end{aligned}$$

Given  $k$ , sampling from the posterior is thus straightforward. We now propose a way to calibrate the hyperparameters  $a, b, \mu, m$  and  $M_0$ .

We first propose a calibration for  $a, b, m, \mu$  and  $\lambda$ . We choose  $m$  to be the empirical mean of the  $Y_i$ . We then chose  $a$  and  $b$  such that the prior on  $\sigma$  has a first order moment and  $E_\pi(\sigma^2)$  is of the same order as the empirical variance of the data  $Y^n$  denoted  $\hat{\sigma}_y^2$ . We choose  $a = \hat{\sigma}_y^2 + 1$  and  $b = \hat{\sigma}_y^4$ . We want the prior on  $\omega$  to be flat enough to recover large variations from the mean  $m$ . This is done by choosing the hyperparameter  $\mu$  small. We also want the prior on  $k$  to be flat to allow large values of  $k$  even for small samples sizes. It seems that  $\mu$  and  $\lambda$  do not have a great influence on the results when performing our test on simulated data. We thus fix  $\mu = 10^{-1}$  and  $\lambda = 10^{-1}$ .

Given these choices for  $a, b, m, \lambda$  and  $\mu$ , we calibrate  $M_0$  the constant in  $\tau_n^k$ . The choice of  $M_0$  is critical for small sample sizes. Given that flats parts of the functions are the most difficult to detect, especially when  $k$  is large, we let  $M_0$  depend on  $k$  and calibrate it on simulated data from the completely flat function  $f = 0$  in order to get an upper bound for the type I error for finite sample sizes.

We denote  $Y_0^n$  data generated from model (3.1) with  $f = 0$  and noise level  $\sigma$ . For all  $k$  we denote  $Z(Y_0^n, k)$  the posterior median of  $H(\omega, k)$  given  $k$  i.e.

$$Z(Y^n, k) = \inf \{z, \Pi\{H(\omega, k) > z | Y_0^n, k\} \geq 1/2\}.$$

We then compute for each  $k$ ,  $M_t(k)$  the  $1 - t$  quantile of  $Z(Y^n, k)$ . It is natural to assume that the constant  $M_0$  should be proportional to the noise level  $\sigma$ . Hence a calibration for  $M_0$

$$M_0 = M_t(k)\sigma^{-1} \left\{ \frac{n}{k \log(n)} \right\}^{1/2}.$$

For each  $k$  sampled from the posterior, we use simple Monte-Carlo approximation for  $M_t(k)$ , based on  $10^3$  samples under the posterior to approximate  $Z(Y_0^n, k)$  and  $10^3$  replications of  $Y_0^n$  to approximate  $M_t(k)$ .

### 3.3 Simulated Examples

In this section we run our testing procedure on simulated data to study the behaviour of our test for finite sample sizes. We choose the prior distribution and

calibrate the hyperparameters as exposed in section 3.2.3. We consider the following nine functions adapted from Baraud et al. (2003) and plot in Figure 3.2.

$$\begin{aligned}
 f_1(x) &= -15(x-0.5)^3 \mathbb{I}_{x \leq 1/2} - 0.3(x-0.5) + e^{-250(x-0.25)^2} \\
 f_2(x) &= 0.15x \\
 f_3(x) &= 0.2e^{-50(x-0.5)^2} \\
 f_4(x) &= -0.5 \cos(6\pi x) \\
 f_5(x) &= -0.2x + f_3(x) \\
 f_6(x) &= -0.2x + f_4(x) \\
 f_7(x) &= -(1+x) + 0.45e^{-50(x-0.5)^2} \\
 f_8(x) &= -0.5x^2 \\
 f_9(x) &= 0
 \end{aligned} \tag{3.12}$$

The functions  $f_1$  to  $f_6$  are clearly not in  $\mathcal{F}$ . The function  $f_7$  has a small bump

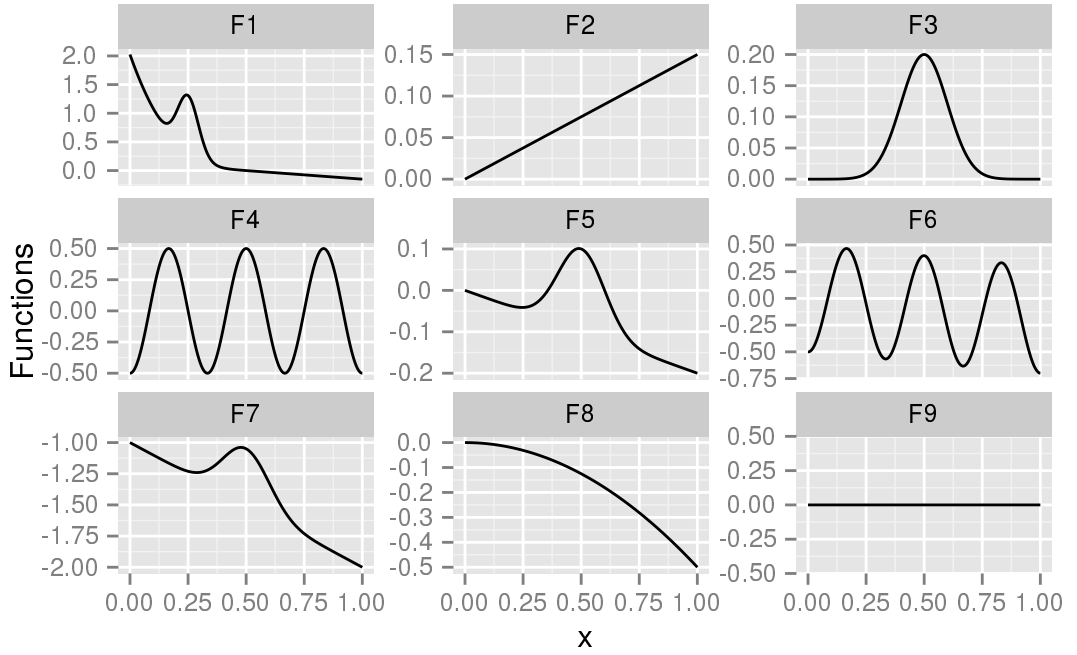


Figure 3.2: Regression functions used in the simulated example.

around  $x = 0.5$  which can be seen as a local departure from monotonicity. This function is thus expected to be difficult to detect for small datasets given our parametrization. The function  $f_9$  is a completely flat function.

Table 3.1: Percentage of rejection for the simulated examples

	$f_0$	$\sigma^2$	Barraud et al. $n = 100$	Akakpo et al. $n = 100$	Bayes Test, $n :$				
					100	250	500	1000	2500
$H_1$	$f_1$	0.01	99	99	97	100	100	100	100
	$f_2$	0.01	99	100	95	100	100	100	100
	$f_3$	0.01	99	98	100	100	100	100	100
	$f_4$	0.01	100	99	100	100	100	100	100
	$f_5$	0.004	99	99	100	100	100	100	100
	$f_6$	0.006	98	99	100	100	100	100	100
	$f_7$	0.01	76	68	97	100	100	100	100
$H_0$	$f_8$	0.01	-	-	2	0	0	0	0
	$f_9$	0.01	-	-	2	3	2	2	0

For several values of  $n$ , we generate  $N = 500$  replication of the data  $Y^n = \{y_i, i = 1, \dots, n\}$  from model (3.1). For each replication we draw  $K = 5.10^3$  iterations from the posterior distribution using a Hasting-Metropolis sampler with a compound Geometric proposal. More precisely, if  $k_{i-1}$  the state of our Markov chain at the step  $i$ , we propose

$$k_i^p = k_{i-1} + p_i$$

where  $p_i$  is such that

$$|p_i| \sim \text{Geom}(0.3) + 1$$

$$P(p_i < 0) = P(p_i > 0) = \frac{1}{2}$$

Given  $k$  we draw directly  $\sigma^2$  and  $\omega$  from the marginal posteriors. We then approximate  $\pi \{H(\omega, k) > \tau_n^k | Y^n\}$  by the standard Monte Carlo estimate

$$\hat{\pi} \{H(\omega, k) > \tau_n^k | Y^n\} = \frac{1}{K} \sum_{i=1}^K \mathbb{I} \{H(\omega^i, k^i) > \tau_n^{k^i}\}$$

and reject the null if  $\hat{\pi} \{H(\omega, k) > \tau_n^k | Y^n\} > 1/2$ . The results are given in table 3.1.

For all the considered functions, the computational time is reasonable even for large values of  $n$ . For instance, for  $f_1$ , we require less than 45 seconds to perform the test for  $n = 2500$  using a simple Python script available on the author's webpage. For the models with regression function  $f_1$  to  $f_7$ , we choose the same residuals

variance as in Baraud et al. (2003), for the last two functions, we choose a variance of 0.01 which is of the same order. We observe that for the regression functions  $f_1$  to  $f_7$ , the test perform well and reject monotonicity for almost all tested samples even when  $n$  is small. The results obtained for  $n = 100$  are comparable with those obtained in Akakpo et al. (2014) and Baraud et al. (2003). For  $f_7$ , our test outperforms the frequentist procedures. Although the Bayesian approach does not fix a level for the test, it appears that with our hyperparameter calibration, the Type 1 error is indeed less or equal to the level of 5% fixed for the frequentist tests.

### 3.4 Application to Global Warming data

We consider the Global Warming dataset provided by Jones et al. (2011) plotted in Figure 3.4. It contains the annual temperatures anomalies from 1850 to 2010, expressed in degrees Celcius. Temperature anomaly is the departure from a long-term average, here the 1961-1990 mean. The data are gathered from both land and sea meteorological stations and corrected for non climatic error. In the literature, this dataset has been used to illustrate some isotonic regression techniques in Wu et al. (2001) and Zhao and Woodroffe (2012) where they use frequentist estimation procedures under monotonicity constraint. Alvarez and Dey (2009) show, using a Bayesian monotonic change point method, that there is a positive trend, and that this trend tends to increase of about  $.3^\circ C$  in the global annual temperature between 1958 and 2000. Álvarez and Yohai (2012) show that the phenomenon of global warming is due to a steady increase trend phenomenon using a isotonic estimation methods. In our model, that would mean that the regression function  $f$  should be positive increasing and convex. In all these papers the data is supposed to be a sequence of independent and identically distributed random variables. This assumption is questionable (see Fomby and Vogelsang (2002)), but considering annual temperature anomalies should reduce the serial correlation. Similarly to these authors, we make the same assumption of independence. Our aim is to test if the hypothesis of increasing temperature anomaly is realistic, given the amount of information, using the method described in section 3.1.1. In particular, we choose the prior and the hyperparameters based on the rules described in section 3.2.

We perform our test on this dataset (more precisely on minus the temperature anomalies to test for monotone increasing trend), choosing the hyperparameters as in section 3.2.3. We run the MCMC sampler described above for  $K = 10^5$  in order to compute Monte Carlo estimate of  $\delta_n^\pi$ . We obtained

$$\hat{\pi}(H(\omega, k) > \tau_n^k | Y^n) = 0.98$$

and thus the hypothesis of monotony is ruled out by our procedure. We conclude that applying a shape constraint regression techniques on the trend of this dataset

can deteriorate the estimation results.

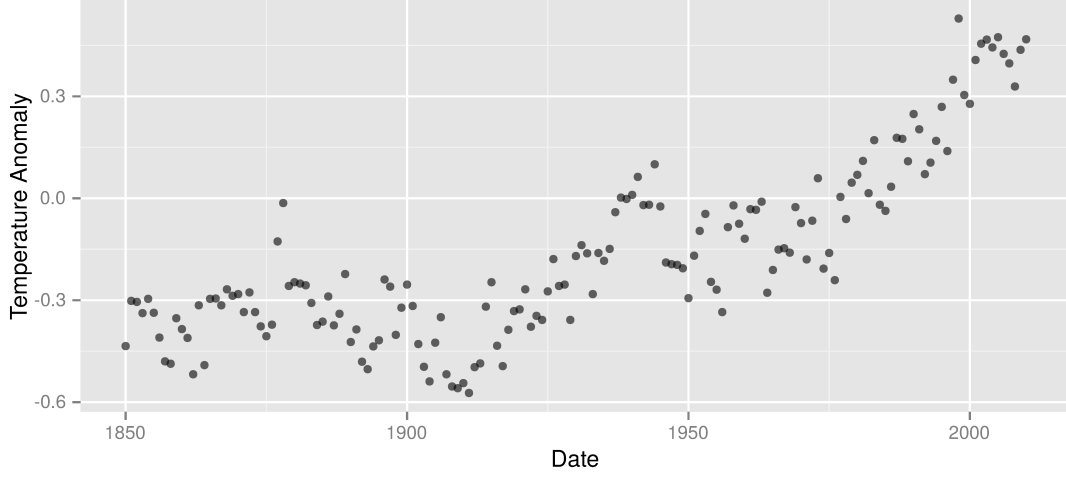


Figure 3.3: Plot of the Global Warming data

### 3.5 Proof of Theorem 3.1

Throughout the proof, we will denote by  $C$  generic constants. Given that we consider  $K$  to be fixed, we will write  $\mathcal{F}$  instead of  $\mathcal{F}(K)$  to lighten notations. In order to prove Theorem 4.1 we need some concentration results of the posterior around the true regression function. The following Lemma provides a posterior concentration rate when  $f_0$  is either in  $\mathcal{F}$  or in  $\mathcal{H}(\alpha, L)$ . The proof is given in Section 3.6 and is derived from Ghosal and van der Vaart (2007). Some adaptive results are known for the Gaussian regression under some regularity assumptions, the monotone case has not been studied and thus this Lemma has an interest in its own.

Let  $d_n(\cdot, \cdot)$  be defined as

$$d_n(f, g)^2 = n^{-1} \sum_{i=1}^n \{f(i/n) - g(i/n)\}^2$$

and denote  $P_0^n$  the distribution of the  $Y_i$  when  $f = f_0$  in (3.1).

**Lemma 3.1.** *Let  $f_0$  be either in  $\mathcal{F}$  or in  $\mathcal{H}(\alpha, L)$ , and let  $\pi$  be defined as in Theorem 4.1. Thus*

$$\mathbb{E}_{P_0^n} [\Pi\{d_n(f_{\omega,k} - f_0)^2 + (\sigma - \sigma_0)^2 \geq \epsilon_n^2 | Y^n\}] \rightarrow 0$$

where  $\epsilon_n = \epsilon_n(\mathcal{F}) = C_K \{n/\log(n)\}^{-1/4}$  if  $f_0 \in \mathcal{F}$ ,  $C_K$  depending only on  $K$  and  $\Pi$  and  $\epsilon_n = \epsilon_n(\alpha) = C_L \{n/\log(n)\}^{-\alpha/(2\alpha+1)}$  if  $f_0 \in \mathcal{H}(\alpha, L)$ ,  $C_L$  depending only on  $L$  and  $\Pi$ .

The proof of this lemma is postponed to Section 3.6. Given this result, we get the following Lemma that enable us to derive consistency and an upper bound on the separation rate.

**Lemma 3.2.** *Let  $M$  be a positive constant and  $\rho_n(\alpha) = M \{n/\log(n)\}^{-\alpha/(2\alpha+1)}$ . Let  $\Pi$  be as in Theorem 4.1 and  $\omega_0$  be the minimizer of the Kulback-Leibler divergence  $KL(f_{\omega,k}, f_0)$ . Then there exists a constant  $A > 0$  such that*

$$P_0^n \left\{ \Pi \left( \max_i |\omega_i - \omega_i^0| \geq A \xi_n^k |Y^n \right) \leq \frac{\gamma_1}{\gamma_0 + \gamma_1} \right\} \rightarrow 1. \quad (3.13)$$

where  $\xi_n^k = [\{k \log(n)\}/n]^{1/2}$  for all fixed positive  $\gamma_0$  and  $\gamma_1$ .

The proof of this lemma is postponed to Section 3.6. Given the preceding results, we derive (3.9).

We first prove consistency under  $H_0$ . Let  $f_0 \in \mathcal{F}$  then

$$H(\omega, k) \leq 2 \max_i |\omega_i - \omega_i^0|$$

and thus

$$P_0^n \left[ \Pi \{H(\omega, k) \geq \tau_n^k |Y^n\} < \frac{\gamma_1}{\gamma_0 + \gamma_1} \right] \rightarrow 1$$

as soon as  $\tau_n^k \geq 2A \xi_n^k$ , which gives the consistency under  $H_0$  given Lemma 3.2.

We now prove consistency under  $H_1$ . Let  $f_0 \notin \mathcal{F}$  and  $f_0 \in \mathcal{H}(\alpha, L)$  we have

$$H(\omega, k) \geq H(\omega^0, k) - 2 \max_i |\omega_i - \omega_i^0| \quad (3.14)$$

Assume that  $\rho_n(\alpha) < d_\infty(f_0, \mathcal{F})$ , we derive a lower bound for  $H(\omega^0, k)$ . Let  $g^*$  be the monotone non increasing piecewise constant function on the partition  $\{[0, 1/k), \dots, [(k-1)/k, 1)\}$ , with for  $1 \leq i \leq k$ ,  $g_i^* = \min_{j \leq i} \omega_j^0$ . Given that  $d_\infty(f_{\omega^0, k}, \mathcal{F}) = \inf_{g \in \mathcal{F}} d_\infty(f_{\omega^0, k}, g)$  we get

$$d_\infty(f_{\omega^0, k}, \mathcal{F}) \leq d_\infty(f_{\omega^0, k}, g^*) \leq H(\omega^0, k)$$

And therefore, given that  $d_\infty(f_0, \mathcal{F}) \leq d_\infty(f_{\omega^0, k}, \mathcal{F}) + d_\infty(f_{\omega^0, k}, f_0)$

$$\Pi \{H(\omega, k) < \tau_n^k |Y^n\} \leq \Pi \left\{ \max_i |\omega_i - \omega_i^0| \geq \frac{\rho_n(\alpha) - d_\infty(f_{\omega^0, k}, f_0) - C \tau_n^k}{4} |Y^n \right\}$$

The following Lemma states that for  $K_0$  a fixed positive constant, the posterior probability of  $k$  being greater that  $K_0 n \rho_n(\alpha)^2 / \log(n)$  is less than a  $o_{P_0^n}(1)$ .

**Lemma 3.3.** *Let  $k_n = n\epsilon_n^2/\log(n)$  if  $L(k) = \log(k)$  and  $k_n = n\epsilon_n^2$  if  $L(k) = 1$  where  $\epsilon_n$  is either  $\epsilon_n(\mathcal{F})$  if  $f_0 \in \mathcal{F}$  or  $\epsilon_n(\alpha)$  if  $f_0 \in \mathcal{H}(\alpha, L)$ . For  $C_1$  a positive constant that may depend on  $K$  or  $L$ , let  $\mathcal{K}_n = \{k \leq C_1 k_n\}$ . If  $\Pi$  is defined as in Theorem 4.1 then*

$$\Pi(\mathcal{K}_n^c | Y^n) \leq o_{P_0^n}(1) \quad (3.15)$$

The proof is postponed to Section 3.6

For  $k \in \mathcal{K}_n$  and  $M$  large enough we have  $\rho_n(\alpha)/4 > \tau_n^k$ . Denoting  $B_n = \{d_n(f_{\omega,k}, f_0)^2 + |\sigma_0 - \sigma|^2 \leq \epsilon_n^2\}$ , Lemma 3.1 gives

$$\Pi(B_n^c | Y_n) = o_{P_0^n}(1).$$

On the set  $B_n \cap \mathcal{K}_n$  we have for  $M$ , the constant in  $\rho(\alpha)$  large enough  $\rho_n(\alpha)/4 \geq d_\infty(f_{\omega^0,k}, f_0)$

$$\Pi\{H(\omega, k) < \tau_n^k | Y_n\} \leq \Pi\left[\left\{\max_i |\omega_i - \omega_i^0| \geq \rho_n(\alpha)/8\right\} \cap \{B_n\} | Y^n\right] + o_{P_0^n}(1).$$

Given (3.13), we get that for all  $f_0$  such that  $d_n(f_0, \mathcal{F}) > \rho_n(\alpha)$

$$P_0^n \left[ \Pi\{H(\omega, k) < \tau_n^k | Y_n\} < \frac{\gamma_0}{\gamma_0 + \gamma_1} \right] \rightarrow 1$$

which ends the proof.

## 3.6 Proof of Lemmas 3.1, 3.2 and 3.3

### 3.6.1 Proof of Lemma 3.1

In this section we prove that the posterior concentrates around  $f_0, \sigma_0$  at the rate  $(n/\log(n))^{-1/4}$  if  $f_0 \in \mathcal{F}$  and  $(n/\log(n))^{-\alpha/(2\alpha+1)}$  if  $f_0 \in \mathcal{H}(\alpha, L)$ . To do so we follow the approach of Ghosal and van der Vaart (2007). Throughout the proof,  $C$  will denote a generic constant.

Let  $KL(f, g) = \int f \log(f/g)$  be the Kullback-Leibler divergence between the two probability densities  $f$  and  $g$ . We define  $V(f, g) = \int (\log(f/g) - KL(f, g))^2 f$ . We denote  $p_i(\omega, \sigma, k)$  the probability density with respect to the Lebesgue measure of  $Y_i = f_{\omega,k} + \epsilon_i$  when  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$  and  $p_{i,0}$  the true density of  $Y_i$ , i.e. when  $f = f_0$ . We only consider the case where  $f \in \mathcal{F}$ , a similar proof holds when  $f \in \mathcal{H}(\alpha, L)$ . We define

$$B_n(\epsilon) = \left\{ \sum_{i=1}^n KL\{p_i(\omega, \sigma, k), p_{i,0}\} \leq n\epsilon^2, \sum_{i=1}^n V\{p_i(\omega, \sigma, k), p_{i,0}\} \leq n\epsilon^2 \right\}$$

Here  $p(\omega, \sigma, k)$  and  $p_0$  are Gaussian distributions, we can easily compute

$$\begin{aligned} KL\{p_i(\omega, \sigma, k), p_{i,0}\} &= \frac{1}{2} \log \left( \frac{\sigma^2}{\sigma_0^2} \right) - \frac{1}{2} \left( 1 - \frac{\sigma_0^2}{\sigma^2} \right) + \frac{1}{2} \frac{\{f_{\omega,k}(x_i) - f_0(x_i)\}^2}{\sigma^2} \\ V\{p_i(\omega, \sigma, k), p_{i,0}\} &= \frac{1}{2} \left( 1 - \frac{\sigma_0^2}{\sigma^2} \right)^2 + \left[ \frac{\sigma_0^2}{\sigma^2} \{f_{\omega,k}(x_i) - f_0(x_i)\} \right]^2 \end{aligned}$$

We have  $B_n(\epsilon_n) \supset \{d_n^2(f_{\omega,k}, f_0) \leq C\epsilon_n^2, |\sigma^2 - \sigma_0^2|^2 \leq C\epsilon_n^2\}$ .

For  $f_0 \in \mathcal{F}$ , denoting  $\omega_j^0 = n_j^{-1} \sum_{x_i \in I_j} f_0(x_i)$  and  $\underline{x}_j = \inf(I_j)$ ,  $\overline{x}_j = \sup(I_j)$  we have

$$d_n^2(f_{\omega,k}, f_0) = d_n^2(f_0, f_{\omega^0,k}) + d_n^2(f_{\omega,k}, f_{\omega^0,k})$$

and

$$\begin{aligned} d_n^2(f_0, f_{\omega^0,k}) &= \frac{1}{n} \sum_{j=1}^k \sum_{x_i \in I_j} \{f_0(x_i) - f_{\omega^0,k}\}^2 \\ &\leq \frac{1}{n} \sum_{j=1}^k n_j \{f_0(\underline{x}_j) - f_0(\overline{x}_j)\}^2 \\ &\leq \frac{C}{k} \left[ \sum_{j=1}^k \{f_0(\underline{x}_j) - f_0(\overline{x}_j)\} \right]^2 \leq \frac{C \|f_0\|_\infty^2}{k}. \end{aligned}$$

Denoting  $k_n = C \lceil \|f_0\|_\infty^2 \{n / \log(n)\}^{1/2} \rceil$  we deduce that  $B_n(\epsilon_n) \supset \{k = k_n, \|\omega - \omega^0\|_{k_n}^2 \leq \epsilon_n^2, |\sigma^2 - \sigma_0^2| \leq \epsilon_n^2\}$  where  $\|\cdot\|_k$  is the standard Euclidean norm in  $\mathbb{R}^k$  i.e. for  $a = (a_1, \dots, a_k) \in \mathbb{R}^k$

$$\|a\|_k^2 = k^{-1} \sum_{i=1}^k a_i^2.$$

We deduce that for a fixed positive constant  $C_0$  that depends on  $\|f_0\|_\infty$ ,

$$\pi\{B_n(\epsilon_n)\} \gtrsim \left( C \inf_{x \in [0,1]} [g\{f_0(x)\}] \epsilon_n \right)^{k_n} \pi_\sigma(\sigma_0^2) \epsilon_n^2 \pi(k = k_n) \geq e^{-C_0 n \epsilon_n^2}. \quad (3.16)$$

To end the proof of Lemma 3.1, the standard approach of Ghosal and van der Vaart (2007) requires the existence of an exponentially consistent sequence of tests. Their Theorem 4 suited for independent observations relies on the fact that the set  $\{d_n(f_{\omega,k}, f_0)^2 + (\sigma - \sigma_0)^2 \geq \epsilon_n^2\}$  can be covered with Hellinger balls. Because of the unknown variance, this cannot be done here, we thus use an alternative approach

and to construct tests, and then apply Theorem 3 from Ghosal and van der Vaart (2007).

Consider the sets  $\mathcal{F}_j^k = \{f_{\omega,k}, \sigma; (j\epsilon_n)^2 \leq d_n(f_{\omega,k}, f_0)^2 + (\sigma - \sigma_0)^2 \leq ((j+1)\epsilon_n)^2\}$ . There exists a constant  $C > 0$  such that

$$\mathcal{F}_j^k \subset \{||\omega - \omega^0||_k \leq Cj\epsilon_n, |\sigma - \sigma_0| \leq Cj\epsilon_n\}. \quad (3.17)$$

To apply Theorem 3 of Ghosal and van der Vaart (2007), we construct tests following Choi and Schervish (2007).

For  $|\sigma - \sigma_0| \leq \sigma_0/2$ . Simple algebra leads to an equivalence between  $(d_n(f, f')^2 + (\sigma - \sigma')^2)^{1/2}$  and the Hellinger metric so that we can apply Lemma 2 of Ghosal and van der Vaart (2007). Equation (3.17) implies that for all  $\xi > 0$  there exist a  $\xi\epsilon_n$  net of  $\mathcal{F}_j^k$  containing less than  $(Cj/\xi)^k$ . We then have a test  $\Psi_1$  such that

$$E_0^n(\Psi_1) \leq e^{-Cj^2n\epsilon_n^2}, \quad \sup_{\mathcal{F}_j^k \cap \{|\sigma - \sigma_0| \leq \sigma_0/2\}} E_{f,\sigma}(1 - \Psi_1) \leq e^{-Cj^2n\epsilon_n^2}.$$

For  $\sigma > 3\sigma_0/2$  we consider the test  $\Psi_2$  defined as

$$\Psi_2 = \mathbb{I} \left\{ \sum_{i=1}^n \left( \frac{Y_i - f_0(x_i)}{\sigma_0} \right)^2 > nc_1 \right\},$$

for a suitably chosen constant  $c_1 > 0$ . Chernoff bound gives

$$E_0^n(\Psi_2) \leq e^{-Cn}.$$

If  $\sigma > 3\sigma_0/2$  and  $(f, \sigma) \in \mathcal{F}_j^k$ , thus  $j > j_0/\epsilon_n$  for some  $j_0 > 0$ . If  $Y_i = f(x_i) + \sigma\epsilon_i$  where  $\epsilon_i \sim \mathcal{N}(0, 1)$  then  $\sum_{i=1}^n ((Y_i - f_0(x_i))/\sigma_0)^2$  follow a non central  $\chi_n^2$  distribution with non centrality parameter  $\sum_{i=1}^n (f(x_i) - f_0(x_i))^2/\sigma^2 > 0$ . Thus setting  $W \sim \chi_n^2$

$$E_{f,\sigma}(1 - \Psi_2) = P_{f,\sigma} \left[ \frac{\sigma^2}{\sigma_0^2} \sum_{i=1}^n \left\{ \frac{Y_i - f_0(x_i)}{\sigma} \right\}^2 \leq nc_1 \right] \leq \text{pr} \left( W \leq \frac{4}{9}c_1n\frac{\sigma_0^2}{\sigma} \right).$$

Chernoff bound gives

$$E_{f,\sigma}(1 - \Psi_2) \leq e^{-C_2n}.$$

Recall that we can construct a  $\xi\epsilon$ -net for  $\mathcal{F}_j^k$  with less than  $(Cj/\xi)^k$  points. For  $\sigma < \sigma_0/2$  we consider the test  $\Psi_3^t$  associated to  $f^t \in \mathcal{F}_j^k$  a point in the  $\xi\epsilon_n$  net and some suitably chosen  $0 < c_2 < 1$  defined as

$$\Psi_3^t = \mathbb{I} \left[ \sum_{i=1}^n \left\{ \frac{Y_i - f^t(x_i)}{\sigma_0} \right\}^2 \leq c_2n \right].$$

As before, given that under  $P_{f_0, \sigma_0}$ ,  $\sum_{i=1}^n [\{Y_i - f^t(x_i)\}/\sigma_0]^2$  follows a non central  $\chi_n^2$  distribution

$$E_0^n(\Psi_3^t) = P_0 \left[ \sum_{i=1}^n \left\{ \frac{Y_i - f^t(x_i)}{\sigma_0} \right\}^2 \leq c_2 n \right] \leq \text{pr}(W \leq c_2 n).$$

Given that the moment generating function of a non central  $\chi_n^2$  distribution with non centrality parameter  $\Delta$  at point  $s$  is known to be  $(1-2s)^{n/2} \exp\{s\Delta^2/(1-2s)\}$ , we have for all  $f, \sigma \in \mathcal{F}_j^k \cap \{\sigma < \sigma_0/2\}$  such that  $d_n(f^t, f) \leq \xi \epsilon_n$

$$\begin{aligned} P_{f, \sigma} \left[ \frac{\sigma^2}{\sigma_0^2} \sum_{i=1}^n \left\{ \frac{Y_i - f^t(x_i)}{\sigma} \right\}^2 \geq c_2 n \right] \\ \leq \exp \left[ \frac{n}{2} \left\{ -\log(1-2s) + \frac{1}{\sigma^2} \frac{2s}{1-2s} d_n(f, f^t)^2 - 2sc_2 \frac{\sigma_0^2}{\sigma^2} \right\} \right]. \end{aligned}$$

For  $s$  small enough we have

$$\frac{2s}{1-2s} d_n(f, f^t)^2 \leq 4s d_n(f, f^t)^2 \leq 4s \xi^2 \epsilon_n^2 \leq 2sc_2 \frac{\sigma_0^2}{\sigma^2}.$$

Which in turns gives for  $c'_2 > 0$  a fixed constant

$$E_{f, \sigma}(1 - \Psi_3^t) \leq e^{-nc'_2}.$$

Taking  $\Psi_3 = \max_t \Psi_3^t$  we get a test such that

$$E_0^n(\Psi_3) = o(1); \quad \sup_{\mathcal{F}_n^j \cap \{\sigma \leq \sigma_0/2\}} E_{f, \sigma}(1 - \Psi_3) \leq e^{-Cj^2 n \epsilon_n^2}.$$

We conclude the proof by taking  $\phi_n = \max\{\Psi_1, \Psi_2, \Psi_3\}$  as an exponentially consistent sequence of tests and applying Theorem 3 of Ghosal and van der Vaart (2007).

### 3.6.2 Proof of lemma 3.2

Let  $f_0$  either belong to  $\mathcal{F}$  or to  $\mathcal{H}(\alpha, L)$  and  $\epsilon_n$  represent either  $\epsilon_n(\mathcal{F})$  if  $f_0 \in \mathcal{F}$  or  $\epsilon_n(\alpha)$  if  $f_0 \in \mathcal{H}(\alpha, L)$ . We denote  $A_n = \{(\omega, \sigma, k), d_n(f_{\omega, k}, f_0)^2 + |\sigma - \sigma_0|^2 \leq \epsilon_n^2\}$  with  $\epsilon_n$  as in Lemma 3.1. Thus  $\pi(A_n^c | Y_n) = o_{P_0^n}(1)$ . We now derive an upper bound for  $\pi(\max_j |\omega_j - \omega_j^0| \geq A \xi_n^k | Y_n, A_n)$ . To do so, we look at the following

decomposition for all  $k_n \in \mathbb{N}$ ,

$$\begin{aligned} & \pi(\max_j |\omega_j - \omega_j^0| \geq A\xi_n^k |Y_n, A_n) \leq \\ & \sum_{k \leq k_n} \pi(k | Y_n, A_n) \sum_{j=1}^k \int \pi(|\omega_j - \omega_j^0| \geq C\xi_n^k | Y_n, A_n, k, \sigma) d\pi(\sigma | Y_n, A_n, k) + \pi(k > k_n | Y_n). \end{aligned} \quad (3.18)$$

Given Lemma 3.3 we have, choosing  $k_n = C_1 n \epsilon_n^2$  a constant  $C_1$  as in Lemma 3.3,

$$\pi(k > k_n | Y_n) = o_{P_0^n}(1)$$

We now find an upper bound uniformly in  $\sigma$  over  $A_n$  for  $\pi(|\omega_j - \omega_j^0| \geq A\xi_n^k | Y_n, A_n, k, \sigma)$ . We first denote  $I_l(\omega_j^0, \sigma_0) = \{l\sigma_0\xi_n^k \leq |\omega_j - \omega_j^0| \leq (l+1)\sigma_0\xi_n^k\}$ . We have for  $l_0 \leq A$

$$\Pi(|\omega_j - \omega_j^0| \geq A\xi_n^k | Y_n, A_n, k, \sigma) \leq \sum_{l \geq l_0} \Pi\{I_l(\omega_j^0, \sigma_0) | Y_n, A_n, k, \sigma\}.$$

We then write

$$\Pi\{I_l(\omega_j^0, \sigma_0) | Y_n, A_n, k, \sigma\} = \frac{\int_{I_l(\omega_j^0, \sigma_0)} e^{l_n^\sigma(\omega) - l_n^{\sigma_0}(\omega^0)} d\Pi(\omega)}{\int e^{l_n^\sigma(\omega) - l_n^{\sigma_0}(\omega^0)} d\Pi(\omega)},$$

where  $l_n^\sigma(\omega) = -n \log(\sigma^2)/2 - \frac{1}{2} \sum_{i=1}^n \{Y_i - f_{\omega, k}(x_i)\}^2 / \sigma^2$ . Standard algebra leads to

$$l_n^\sigma(\omega) - l_n^{\sigma_0}(\omega^0) = -\frac{1}{2} \sum_{j=1}^k \frac{(\omega_j - \omega_j^0)^2}{\sigma^2} + \sum_{x_i \in I_j} \epsilon_i \frac{\sigma_0}{\sigma^2} (\omega_j - \omega_j^0) + \Delta(\epsilon, \sigma, f_0, k),$$

where  $\Delta(\epsilon, \sigma, f_0, k)$  does not depend on  $\omega$  and  $\epsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, 1)$  under  $p_0^n$ . We thus deduce

$$\begin{aligned} & \Pi\{I_l(\omega_j^0, \sigma_0) | Y_n, A_n, k, \sigma\} = \\ & \frac{\int_{I_l(\omega_j^0, \sigma_0)} \exp \left\{ -\frac{1}{2} n_j \frac{(\omega_j - \omega_j^0)^2}{\sigma^2} + \sum_{x_i \in I_j} (\epsilon_i) \frac{\sigma_0}{\sigma^2} (\omega_j - \omega_j^0) \right\} d\Pi(\omega)}{\int \exp \left\{ -\frac{1}{2} n_j \frac{(\omega_j - \omega_j^0)^2}{\sigma^2} + \sum_{x_i \in I_j} (\epsilon_i) \frac{\sigma_0}{\sigma^2} (\omega_j - \omega_j^0) \right\} d\Pi(\omega)} = \frac{N_{n,j,l}^k(\sigma)}{D_{n,j}^k(\sigma)} \end{aligned}$$

We now prove that on a set  $\mathcal{E}$  such that  $P_0^n(\mathcal{E}) = 1 + o(1)$  we have for  $(\epsilon_i) \in \mathcal{E}$ , We have an upper bound for  $N_{n,j}^k / D_{n,j}^k$  uniformly in  $\sigma \in A_n$  for all  $k \leq k_n$ .

Let  $\mathcal{E} = \left\{ \cap_{k \leq k_n} \cap_{j=1}^k \left\{ \left| \sum_{x_i \in I_j} \epsilon_i \right| \leq c_e \sqrt{n_j \log(n)} \right\} \right\}$  for some constant absolute constant  $c_e$  large enough. We compute

$$\text{pr}(\mathcal{E}^c) \leq 2 \sum_{k=2}^{k_n} \sum_{j=1}^k \text{pr} \left( \sum_{x_i \in I_j} \epsilon_i > c_e \sqrt{n_j \log(n)} \right) \leq 2 \frac{k_n^2}{n^{c_e^2}} = o(1).$$

For  $(\epsilon_i) \in \mathcal{E}$  and uniformly in  $\sigma$  over  $A_n$  we compute

$$\begin{aligned} D_{n,j}^k(\sigma) &= \int \exp \left\{ -\frac{n_j}{2\sigma^2} (\omega_j - \omega_j^0)^2 + \frac{\sigma_0}{\sigma^2} (\omega_j - \omega_j^0) \sum_{x_i \in I_j} \epsilon_i \right\} d\pi(\omega_j) \\ &\geq \int_{|\omega_j - \omega_j^0| \leq \sigma_0 c_e \xi_n^k} \exp \left\{ -n_j (\omega_j - \omega_j^0)^2 - 2c_e \frac{\sigma_0}{\sigma^2} n_j |\omega_j - \omega_j^0| \sqrt{\frac{\log(n)}{n_j}} \right\} d\pi(\omega_j) \\ &\geq e^{-3c_e^2 \sigma_0^2 n_j (\xi_n^k)^2 / (2\sigma^2)} \Pi(|\omega_j - \omega_j^0| \leq \sigma_0 c_e \xi_n^k) \end{aligned}$$

Similarly for  $(\epsilon_i) \in \mathcal{E}$  and uniformly in  $\sigma$  over  $A_n$  we have for  $l$  large enough

$$\begin{aligned} N_{n,j,l}^k(\sigma) &\leq \int_{I_l(\omega_j^0, \sigma_0)} \exp \left\{ -\frac{1}{2} n_j |\omega_j - \omega_j^0| \left( \frac{|\omega_j - \omega_j^0|}{\sigma^2} - \frac{\sigma_0}{\sigma^2} c_e \sqrt{\frac{\log(n)}{n_j}} \right) \right\} d\pi(\omega) \\ &\leq e^{-l^2 \sigma_0^2 n_j (\xi_n^k)^2 / (4\sigma^2)} \Pi\{I_l(\omega_j^0, \sigma_0)\}. \end{aligned}$$

We thus have for  $(\epsilon_i)_i \in \mathcal{E}$ ,  $\epsilon > 0$  and  $l$  large enough, together with condition **C2**

$$\begin{aligned} \frac{N_{n,j,l}^k(\sigma)}{D_{n,j}^k(\sigma)} &\leq e^{-\frac{1}{2\sigma^2} \sigma_0^2 n_j (\xi_n^k)^2 (l/2 - 3c_e)} \frac{\Pi\{I_l(\omega_j^0, \sigma_0)\}}{\Pi(|\omega_j - \omega_j^0| \leq \sigma_0 c_e \xi_n^k)} \\ &\leq e^{-n_j (\xi_n^k)^2 l^2 \frac{\sigma_0^2}{8\sigma^2}}, \end{aligned}$$

which in turns gives an upper bound for  $\Pi(|\omega_j - \omega_j^0| \geq A \xi_n^k | Y_n, A_n, k, \sigma)$

$$\Pi(|\omega_j - \omega_j^0| \geq A \xi_n^k | Y_n, A_n, k, \sigma) \leq \frac{1}{2} e^{-l_0 \frac{\sigma_0^2}{8\sigma^2} n_j (\xi_n^k)^2}.$$

We thus deduce for  $C$  an absolute constant

$$\Pi\left(\max_{1 \leq j \leq k} |\omega_j - \omega_j^0| \geq A \xi_n^k | Y_n\right) \leq k_n e^{-l_0 C \log(n)} + o_{P_0^n}(1),$$

which gives choosing  $A$  large enough

$$P_0^n \left\{ \Pi\left(\max_{1 \leq j \leq k} |\omega_j - \omega_j^0| \geq A \xi_n^k | Y_n\right) < \frac{\gamma_1}{\gamma_0 + \gamma_1} \right\} \rightarrow 1.$$

### 3.6.3 Proof of Lemma 3.3

Let be either  $k_n = n\epsilon_n^2/\log(n)$  if  $L(k) = \log(k)$  or  $k_n = n\epsilon_n^2$  if  $L(k) = 1$ . Similarly to before, we have  $\pi(B_n(\epsilon_n)) \geq e^{-n\epsilon_n^2}$ . We define  $N_n$  and  $D_n$  such that

$$\pi(\mathcal{K}_n^c|Y_n) = \frac{\sum_{k \in \mathcal{K}_n^c} \pi(k) \int \frac{p(\omega, \sigma, k)}{p_0}(Y^n) d\Pi(\omega, \sigma)}{\sum_k \pi(k) \int \frac{p(\omega, \sigma, k)}{p_0}(Y^n) d\Pi(\omega, \sigma)} = \frac{N_n}{D_n}$$

Given Lemma 10 of Ghosal and van der Vaart (2007), we have

$$P_0^n \left( D_n \leq e^{-Cn\epsilon_n^2} \right) = o(1)$$

Note also that

$$E_0^n(N_n) = \sum_{k \in \mathcal{K}_n^c} \pi(k) \int \int_{\mathbb{R}^n} \frac{p(\omega, \sigma, k)}{p_0}(Y^n) p_0(Y_n) d\Pi(\omega, \sigma) dY^n = \pi(k \leq k_n) \leq ce^{-C_u k_n L(k_n)}$$

Thus for  $C$  small enough we have

$$\begin{aligned} E_0^n [\Pi(k \in \mathcal{K}_n^c|Y^n)] &= E_0^n \left[ \frac{N_n}{D_n} \mathbb{I}_{D_n > e^{-Cn\epsilon_n^2}} \right] + o(1) \\ &\leq e^{Cn\epsilon_n^2} ce^{-C_u k_n L(k_n)} + o(1) \\ &\leq o(1) \end{aligned}$$

## 3.7 Discussion

In this chapter we propose a Bayesian approach to the problem of testing qualitative hypotheses in a nonparametric framework. More precisely we address the problem of testing monotonicity of a regression function. This problem arise naturally as shape constraint models, and monotonicity in particular, are fairly used in practice. Our approach is particularly interesting as it focuses on a problem where the Bayes Factor seems to give poor results and thus an alternative approach should be considered. The testing procedure proposed in this chapter is a modified version of the Bayes Factor that only reject  $H_0$  when the data gives strong evidence that the function is not monotone. When possible, one can choose a threshold based on prior information on the tolerance level to non monotony. However, this could be difficult in practice, we thus present a way to calibrate our test such that it behave well asymptotically. Interestingly this calibration leads to the optimal separation rate (up to a  $\log(n)$  term) and thus the tolerance induced by our approach, and the fact that we test (3.3),  $H_0^a$  versus  $H_1^a$ , instead of (3.2),

$H_0$  versus  $H_1$ , is of the same order as the classical tests available in the literature. It has the advantage of being very simple to implement even in presence of large datasets. Although we have focused on monotonicity constraints, other types of shape constraints such as convexity or unimodality can be dealt with using this approach. For instance we can test for convexity using piecewise linear functions as submodels  $\mathcal{G}_k$  and test monotonicity of the slope.

## Bibliography

- Akakpo, N., Balabdaoui, F., and Durot, C. (2014). Testing monotonicity via local least concave majorants. *Bernoulli*, 20(2):514–544.
- Alvarez, E. E. and Dey, D. K. (2009). Bayesian isotonic changepoint analysis. *Ann. Inst. Statist. Math.*, 61(2):355–370.
- Álvarez, E. E. and Yohai, V. J. (2012). M-estimators for isotonic regression. *J. Statist. Plann. Inference*, 142(8):2351–2368.
- Baraud, Y., Huet, S., and Laurent, B. (2003). Adaptive tests of qualitative hypotheses. *ESAIM Probab. Stat.*, 7:147–159.
- Baraud, Y., Huet, S., and Laurent, B. (2005). Testing convex hypotheses on the mean of a Gaussian vector. Application to testing qualitative hypotheses on a regression function. *Ann. Statist.*, 33(1):214–257.
- Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression*. John Wiley & Sons, London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics.
- Bornkamp, B. and Ickstadt, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose-response analysis. *Biometrics*, 65(1):198–205.
- Brunner, L. J. and Lo, A. Y. (1989). Bayes methods for a symmetric unimodal density and its mode. *Ann. Statist.*, 17(4):1550–1566.
- Choi, T. and Schervish, M. J. (2007). On posterior consistency in nonparametric regression problems. *J. Multivariate Anal.*, 98(10):1969–1987.
- Fomby, T. B. and Vogelsang, T. J. (2002). The Application of Size-Robust Trend Statistics to Global-Warming Temperature Series. *J. Climate*, 15:117–123.
- Ghosal, S., Sen, A., and van der Vaart, A. W. (2000). Testing monotonicity of regression. *Ann. Statist.*, 28(4):1054–1082.
- Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.*, 35(1):192–223.
- Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berke-*

- ley, Calif., 1983), Wadsworth Statist./Probab. Ser., pages 539–555, Belmont, CA. Wadsworth.
- Hall, P. and Heckman, N. E. (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *Ann. Statist.*, 28(1):20–39.
- Jones, P., Parker, D., Osborn, T., , and Briffa, K. (2011). Global and hemispheric temperature anomalies, land and marine instrumental records.
- Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Stat.*, 12:351–357.
- Mukerjee, H. (1988). Monotone nonparameteric regression. *Ann. Statist.*, 16(2):741–750.
- Neittaanmäki, P., Rossi, T., Majava, K., and Pironneau, O. (2008). Monotonic regression for assesement of trends in environmental quality data.
- Prakasa Rao, B. L. S. (1970). Estimation for distributions with monotone failure rate. *Ann. Math. Statist.*, 41:507–519.
- Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*, volume 319. Citeseer.
- Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester.
- Rousseau, J. (2007). Approximating interval hypothesis:  $p$ -values and Bayes factors. In *Bayesian statistics 8*, Oxford Sci. Publ., pages 417–452. Oxford Univ. Press, Oxford.
- Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Stat.*, 38(1):146–180.
- Salomond, J.-B. (2013). Concentration rate and consistency of the posterior under monotonicity constraints. *ArXiv e-prints*.
- Scott, J. G., Shively, T. S., and Walker, S. G. (2013). Nonparametric Bayesian testing for monotonicity. *ArXiv e-prints*.
- Wu, W. B., Woodroffe, M., and Mentz, G. (2001). Isotonic regression: another look at the changepoint problem. *Biometrika*, 88(3):793–804.
- Zhao, O. and Woodroffe, M. (2012). Estimating a monotone trend. *Statist. Sinica*, 22(1):359–378.



# Chapter 4

## Ill-posed inverse problems

“I may not be as strong as I think, but I know many tricks  
and I have resolution.”

– **Ernest Hemingway**, *The old man and the sea*.

**Co-écrit avec Bartek Knapik**

### Résumé

Nous proposons une méthode générale pour l'étude des problèmes inverses linéaires mal-posés dans un cadre bayésien. S'il existe de nombreux résultats sur les méthodes de régularisation et la vitesse de convergence d'estimateurs classiques, pour l'estimation de fonctions dans un problème inverse mal-posé, les vitesses de concentration d'a posteriori dans le cadre bayésien n'a été que très peu étudié dans ce cadre. De plus ces quelques rares résultats existant ne considèrent que des familles très limitées de lois a priori, en général reposant sur la décomposition en valeurs singulières de l'opérateur considéré. Dans ce chapitre nous proposons des conditions générales sur la loi a priori sous lesquelles l'a posteriori se concentre à une certaine vitesse. Notre approche nous permet de trouver les vitesses de concentration de l'a posteriori pour de nombreux modèles et de larges classes de loi a priori. Cette approche est de plus particulièrement intéressante car elle permet de mieux comprendre le fonctionnement de la loi a posteriori et notamment l'impact de l'opérateur sur l'inférence.

## 4.1 Introduction

Statistical approaches to inverse problems have been initiated in the 1960's and since then many estimation methods have been developed. Inverse problems arise naturally when one only has indirect observations of the object of interest. Mathematically speaking this phenomenon is easily modelled by the introduction of an operator  $K$  such that the observation at hand comes from the model

$$Y^n \sim P_{Kf}^n, \quad (4.1)$$

where  $f$  is the object of interest and is assumed to belong to a parameter space  $\mathcal{F}$ . In many applications the operator  $K$  is assumed to be injective. However, in the most interesting cases its inverse is not continuous, thus the parameter of interest  $f$  cannot be reconstructed by a simple inversion of the operator. Such problems are said to be *ill-posed*. Several methods dealing with the discontinuity of the inverse operator have been proposed in the literature. The most famous one is to conduct the inference while imposing some regularity constraints on the parameter of interest  $f$ . These so-called regularisation methods have been widely studied in the literature both from a theoretical and applied perspective (see Engl et al., 1996, for a review).

Bayesian approach to inverse problems is therefore particularly interesting, as it is well known that putting a prior distribution on the parameter yields a natural regularisation. This property of the Bayesian approach is particularly interesting for model choice, but it has proved also useful in many estimation procedures, as shown in Rousseau and Mengersen (2011) in the case of overfitted mixtures models or to nonparametric models where regularization is necessary as in Castillo (2013) or Salomond (2013) in the semiparametric problem of estimating a monotone density at the boundaries of its support. Here we study the asymptotic behaviour of the posterior distribution under the frequentist assumptions that the data  $Y^n$  are generated from model (4.1) for some true parameter  $f_0$ . In particular we are interested in the rate at which the posterior concentrate around  $f_0$ . Asymptotic properties of the posterior distribution have received a growing interest in the literature. Knapik et al. (2011), Agapiou et al. (2013), and Florens and Simoni (2012) were the first to study posterior concentration rates under conjugate prior in so-called mildly ill-posed setting. These were followed by two papers by Knapik et al. (2013) and Agapiou et al. (2014), studying Bayesian approach to recovery of the initial condition for heat equation and related inverse problems. The paper by Ray (2013) is the first study of the posterior concentration rates in the non-conjugate setting. Considering non-conjugate prior is particularly interesting as it allows some additional flexibility of the model. However, the approach presented in Ray (2013) is only valid for priors that are closely linked to the *singular value decomposition* (SVD) of the operator. Moreover, in Ray (2013) several rate adaptive

priors were considered. It should be noted, however, that some of the bounds on contraction rates obtained in that paper are not optimal. Similar adaptive results, in the conjugate mildly ill-posed setting, using empirical and hierarchical Bayes approach were obtained in Knapik et al. (2012).

There is a rich literature on the problem of deriving posterior concentration rate in the direct problem setting. Since the seminal papers of Ghosal et al. (2000) and Shen and Wasserman (2001), general conditions on the prior distribution for which the posterior concentrates at a certain rate have been derived in various cases. In particular Ghosal and van der Vaart (2007) gives a series of conditions for non independent and identically distributed data. However, such results cannot be applied directly to ill-posed inverse problems and to the authors best knowledge, no equivalent of these results exists in the inverse problem literature. In this work we try to fill this gap. We first assume the existence of the contraction result for the so-called direct problem (that is recovery of  $Kf$ ). Next, we impose additional sufficient conditions on the prior such that the posterior distribution for the parameter of interest  $f$  concentrates at a given rate.

Consider an abstract setting in which the parameter space  $\mathcal{F}$  is an arbitrary metrizable topological vector space and let  $K$  be an injective mapping  $K : \mathcal{F} \ni f \mapsto Kf \in K\mathcal{F}$ . Even if the problem is ill-posed there exist subsets  $\mathcal{S}_n$  of  $K\mathcal{F}$  over which the inverse of the operator can be controlled. For suitably well chosen priors, these sets will capture most of the posterior mass, and we can thus easily derive posterior concentration rate for  $f$  from posterior concentration rate for  $Kf$  by a simple inversion of the operator. More precisely for  $d$  and  $d_K$  some metrics or semi-metrics on  $\mathcal{F}$  and  $K\mathcal{F}$  respectively and  $f_0$  a point in  $\mathcal{F}$ , we want to derive the smallest ball for the metric  $d$  on  $\mathcal{F} \cap \mathcal{S}_n$  that contains  $K^{-1}\{f, d_K(Kf, Kf_0) \leq \epsilon\}$  the image of a ball of  $K(\mathcal{F} \cap \mathcal{S}_n)$  for the metric  $d_K$  by  $K^{-1}$ . This shows in particular that the choice of  $\mathcal{S}_n$  is crucial for our approach.

The rest of the paper is organised as follows: we present the main result in Section 4.2 and a general construction for the sets  $\mathcal{S}_n$  in Section 4.3. We then apply our result for different examples in the white noise and regression setting in Section 4.4.

## 4.2 General Theorem

Assume that the observations  $Y^n$  come from model (4.1) and that  $P_{Kf}^n$  admit densities  $p_{Kf}^n$  relative to a  $\sigma$ -finite measure  $\mu^n$ . To avoid complicated notations, we drop the superscript  $n$  in the rest of the paper. Let  $\mathcal{F}$  and  $K\mathcal{F}$  be metric spaces, and let  $d$  and  $d_K$  denote metrics on both spaces, respectively.

In this section we present the main result of this paper which gives an upper bound on the posterior concentration rate under some general conditions on the

prior. We will call the estimation of  $Kf$  given the observations  $Y$  the *direct problem*, and the estimation  $f$  given  $Y$  the *inverse problem*. The main idea is to control the change of norms between  $d_K$  and  $d$ . If the posterior distribution concentrates around  $Kf_0$  for the metric  $d_K$  at a certain rate in the direct problem, applying the change of norms will give us an upper bound on the posterior concentration rate for the metric  $d$  in the inverse problem. However, since the problem is ill-posed the change of norms cannot be controlled over the whole space  $K\mathcal{F}$ . A way to come around this problem is to only focus on a sequence of sets of high posterior mass for which the change of norm is feasible. More precisely, for a set  $\mathcal{S} \subset \mathcal{F}$ ,  $f_0 \in \mathcal{F}$  and a fixed  $\delta > 0$  we call the quantity

$$\omega(\mathcal{S}, f_0, d, d_K, \delta) := \sup\{d(f, f_0) : f \in \mathcal{S}, d_K(Kf, Kf_0) \leq \delta\}. \quad (4.2)$$

the *modulus of continuity*. We note that in this definition we do not assume  $f_0 \in \mathcal{S}$ . This is thus a local version of the modulus of continuity considered in Donoho and Liu (1991) or Hoffmann et al. (2013). On the one hand, the sets  $\mathcal{S}_n$  need to be big enough to capture most of the posterior mass. On the other hand, one has to be able to control the distance between the elements of  $\mathcal{S}_n$  and  $f_0$ , given the distance between  $Kf$  and  $Kf_0$  is small. Since the operator  $K$  is unbounded, this suggests that the sets  $\mathcal{S}_n$  cannot be too big.

**Theorem 4.1.** *Let  $\epsilon_n \rightarrow 0$  and let  $\Pi$  the prior distribution on  $f$  be such that*

$$\mathbb{E}_0 \Pi(\mathcal{S}_n^c | Y^n) \rightarrow 0, \quad (4.3)$$

*for some sequence of sets  $(\mathcal{S}_n)$ ,  $\mathcal{S}_n \subset \mathcal{F}$ , and*

$$\mathbb{E}_0 \Pi(f : d_K(Kf, Kf_0) \geq M_n \epsilon_n | Y^n) \rightarrow 0,$$

*for any  $M_n \rightarrow \infty$ . Then*

$$\mathbb{E}_0 \Pi(f : d(f, f_0) \geq \omega(\mathcal{S}_n, f_0, d, d_K, M_n \epsilon_n) | Y^n) \rightarrow 0.$$

*Proof.* By (4.3) and the definition of the modulus of continuity

$$\begin{aligned} \Pi(f : d(f, f_0) \geq \omega(\mathcal{S}_n, f_0, d, d_K, M_n \epsilon_n) | Y^n) \\ \leq \Pi(f \in \mathcal{S}_n : d(f, f_0) \geq \omega(\mathcal{S}_n, f_0, d, d_K, M_n \epsilon_n) | Y^n) + \Pi(\mathcal{S}_n^c | Y^n) \\ \leq \Pi(f \in \mathcal{S}_n : d_K(Kf, Kf_0) \geq M_n \epsilon_n | Y^n) + o_P(1). \end{aligned}$$

□

The interpretation of the theorem is the following: given a properly chosen sequence of sets  $\mathcal{S}_n$ , the rate of posterior contraction in the direct problem restricted

to the given sequence can be translated to the rate of posterior contraction in the inverse setting. Here, the choice of  $\mathcal{S}_n$  is crucial as it is the principal component in the control of the change of norm. In particular, the concentration rate  $\epsilon_n$  for the direct problem may not be optimal, and still leads to an optimal concentration rate  $\omega(\mathcal{S}_n, f_0, d, d_K, M_n \epsilon_n)$  for the inverse problem with a well suited choice of  $\mathcal{S}_n$ . As shown in Section 4.4.1.2, this is the case for instance when the posterior distribution of  $Kf$  is very concentrated. We can then choose  $\mathcal{S}_n$  small enough so that the change of norms can be controlled very precisely.

To control the posterior mass of the sets  $\mathcal{S}_n$  we can usually alter the proofs of contraction results for the direct problems. Here we present a standard argument leading to (4.3). Define the usual Kullback–Leibler neighborhoods by

$$B_n(Kf_0, \epsilon) = \left\{ f \in \mathcal{F} : - \int p_{Kf_0} \log \frac{p_{Kf}}{p_{Kf_0}} d\mu \leq n\epsilon^2, \right. \\ \left. \int p_{Kf_0} \left( \log \frac{p_{Kf}}{p_{Kf_0}} \right)^2 d\mu \leq n\epsilon^2, \right\}, \quad (4.4)$$

The following Lemma adapted from Ghosal and van der Vaart (2007) gives general conditions on the prior such that (4.3) is satisfied.

**Lemma 4.1** (Lemma 1 in Ghosal and van der Vaart, 2007). *Let  $\epsilon_n \rightarrow 0$  and let  $(\mathcal{S}_n)$  be a sequence of sets  $\mathcal{S}_n \subset \mathcal{F}$ . If  $\Pi$  is the prior distribution on  $f$  satisfying*

$$\frac{\Pi(\mathcal{S}_n^c)}{\Pi(B_n(Kf_0, \epsilon_n))} \lesssim \exp(-2n\epsilon_n^2),$$

*then*

$$E_0 \Pi(\mathcal{S}_n^c | Y^n) \rightarrow 0.$$

### 4.3 Modulus of continuity

In this section we first present an example of the sequence of sets  $\mathcal{S}_n$ , and later present how the modulus of continuity for this sequence can be computed in two standard inverse problem settings. We now suppose that  $\mathcal{F}$  and  $K\mathcal{F}$  are separable Hilbert spaces, denoted  $(\mathbb{H}_1, \|\cdot\|_{\mathbb{H}_1})$  and  $(\mathbb{H}_2, \|\cdot\|_{\mathbb{H}_2})$  respectively. We note that the sets  $\mathcal{S}_n$  resemble the sets  $\mathcal{P}_n$  considered in Ray (2013).

As already noted, the operator  $K$  restricted to certain subsets of the domain  $\mathbb{H}_1$  might have a finite modulus of continuity defined in (4.2). Clearly, one wants to construct a sequence of sets  $\mathcal{S}_n$  that in a certain sense approaches the full domain  $\mathbb{H}_1$ . This is understood in terms of the remaining prior mass condition in

Theorem 4.1. Moreover, since we do not require  $f_0$  to be in  $\mathcal{S}_n$ , we need to be able to control the distance between  $f_0$  and  $\mathcal{S}_n$ .

A natural guess is to consider finite-dimensional projections of  $\mathbb{H}_1$ . In this section we go beyond this concept. To get some intuition, consider the Fourier basis of  $\mathbb{H}_1$ . The ill-posedness can be then viewed as too big an amplification of the high frequencies through the inverse of the operator  $K$ . Therefore, one wants to control the higher frequencies in the signal, and thus in the parameter  $f$ .

Since  $\mathbb{H}_1$  is a separable Hilbert space, there exist an orthonormal basis  $(e_i)$  and each element  $f \in \mathbb{H}_1$  can be viewed as an element of  $\ell_2$  and

$$\|f\|_{\mathbb{H}_1}^2 = \sum_{i=1}^{\infty} f_i^2.$$

For given sequences  $k_n \rightarrow \infty$  and  $\rho_n \rightarrow 0$ , and a constant  $c \geq 0$  we define

$$\mathcal{S}_n := \left\{ f \in \ell_2 : \sum_{i>k_n} f_i^2 \leq c\rho_n^2 \right\}. \quad (4.5)$$

If the operator  $K$  is compact, then the spectral decomposition of the self-adjoint operator  $K^T K : \mathbb{H}_1 \rightarrow \mathbb{H}_1$  provides a convenient orthonormal basis. In the compact case the operator  $K^T K$  possesses countably many positive eigenvalues  $\kappa_i^2$  and there is a corresponding orthonormal basis  $(e_i)$  of  $\mathbb{H}_1$  of eigenfunctions, and the sequence  $(\tilde{e}_i)$  defined by  $Ke_i = \kappa_i \tilde{e}_i$  forms an orthonormal conjugate basis of the range of  $K$  in  $\mathbb{H}_2$ . Therefore, both  $f$  and  $Kf$  can be associated with sequences in  $\ell_2$ . Since the problem is ill-posed when  $\kappa_i \rightarrow 0$ , we can assume without loss of generality that the sequence  $\kappa_i$  is decreasing.

Let  $k_n$ ,  $\rho_n$ , and  $c$  in the definition of  $\mathcal{S}_n$  be fixed. Then for any  $g \in \mathcal{S}_n$

$$\begin{aligned} \|g\|_{\mathbb{H}_1}^2 &= \sum_{i=1}^{\infty} g_i^2 = \sum_{i \leq k_n} g_i^2 + \sum_{i > k_n} g_i^2 \\ &\leq \sum_{i \leq k_n} g_i^2 + c\rho_n^2 = \sum_{i \leq k_n} \kappa_i^{-2} \kappa_i^2 g_i^2 + c\rho_n^2 \\ &\leq \kappa_{k_n}^{-2} \sum_{i \leq k_n} \kappa_i^2 g_i^2 + c\rho_n^2 \leq \kappa_{k_n}^{-2} \|Kg\|_{\mathbb{H}_2}^2 + c\rho_n^2. \end{aligned}$$

Let  $f_n$  be the projection of  $f_0$  on the first  $k_n$  coordinates, i.e.,  $f_{n,i} = f_{0,i}$  for  $i \leq k_n$  and 0 otherwise. Moreover, we assume that  $f_0$  belongs to some smoothness class described by a decreasing sequence  $(s_i)$ :

$$\|f_0\|_s^2 = \sum_{i=1}^{\infty} s_i^{-2} f_{0,i}^2 < \infty.$$

The usual Sobolev space of regularity  $\beta$  is defined in that way with  $s_i = i^{-\beta}$ . Therefore, we have

$$\|f_n - f_0\|_{\mathbb{H}_1} \leq s_{k_n} \|f_0\|_s, \quad \|Kf_n - Kf_0\|_{\mathbb{H}_2} \leq s_{k_n} \kappa_{k_n} \|f_0\|_s.$$

Using the triangle inequality twice and keeping in mind that  $f - f_n \in \mathcal{S}_n$  we obtain

$$\begin{aligned} \|f - f_0\|_{\mathbb{H}_1} &\leq \|f - f_n\|_{\mathbb{H}_1} + \|f_n - f_0\|_{\mathbb{H}_1} \\ &\leq \kappa_{k_n}^{-1} \|Kf - Kf_n\|_{\mathbb{H}_2} + \sqrt{c}\rho_n + s_{k_n} \|f_0\|_s \\ &\leq \kappa_{k_n}^{-1} (\|Kf - Kf_0\|_{\mathbb{H}_2} + \kappa_{k_n} s_{k_n} \|f_0\|_s) + \sqrt{c}\rho_n + s_{k_n} \|f_0\|_s \\ &= \kappa_{k_n}^{-1} \|Kf - Kf_0\|_{\mathbb{H}_2} + \sqrt{c}\rho_n + 2\|f_0\|_s s_{k_n}. \end{aligned}$$

We then find an upper bound for the modulus of continuity,

$$\omega(\mathcal{S}_n, f_0, \|\cdot\|_{\mathbb{H}_1}, \|\cdot\|_{\mathbb{H}_2}, \delta) \lesssim \kappa_{k_n}^{-1} \delta + \rho_n + s_{k_n}. \quad (4.6)$$

**Remark 1.** If  $c > 0$ , then  $f_0 \in \mathcal{S}_n$  for  $n$  large enough (depending on  $f_0$ ).

## 4.4 Some models

### 4.4.1 White noise

#### 4.4.1.1 Mildly ill-posed problems

Our first example is based on the well-studied infinite-dimensional normal mean model. In the Bayesian context the problem of direct estimation of infinitely many means has been studied, among others, by Zhao (2000); Shen and Wasserman (2001); Belitser and Ghosal (2003); Ghosal and van der Vaart (2007).

We consider the white noise setting, where we observe an infinite sequence  $Y^n = (Y_1, Y_2, \dots)$  satisfying

$$Y_i = \kappa_i f_i + \frac{1}{\sqrt{n}} Z_i, \quad (4.7)$$

where  $C^{-1}i^{-p} \leq \kappa_i \leq Ci^{-p}$  for some  $p \geq 0$  and  $C \geq 1$ , and  $Z_1, Z_2, \dots$  are independent standard normal random variables. Let  $Kf$  denote the sequence  $\kappa_i f_i$ . In this setting  $\mathbb{H}_1 = \mathbb{H}_2 = \ell_2$ , and the  $\ell_2$ -norm is denoted by  $\|\cdot\|$ .

Since the  $\kappa_i$ 's decay polynomially, the problem is *mildly* ill-posed. Such problems are well studied in the frequentist literature, and we refer the reader to Cavalier (2008) for a nice overview. There are also several papers on properties of Bayes procedures for such problems. The first studies of posterior contraction in mildly ill-posed inverse problems were obtained by Knapik et al. (2011) and

Agapiou et al. (2013). Later, Ray (2013) and Knapik et al. (2012) studied adaptive priors leading to the optimal minimax rate of contraction. Similar problem, with a different noise structure, has been studied by Florens and Simoni (2012).

We put a product prior on  $f$  of the form

$$\Pi = \bigotimes_{i=1}^{\infty} N(0, \lambda_i),$$

where  $\lambda_i = i^{-1-2\alpha}$ , for some  $\alpha > 0$ . Furthermore, the true parameter  $f_0$  is assumed to belong to  $S^\beta$  for some  $\beta > 0$ :

$$S^\beta = \left\{ f \in \ell_2 : \|f\|_\beta^2 := \sum f_i^2 i^{2\beta} < \infty \right\}. \quad (4.8)$$

Therefore,  $\|Kf_0\|_{\beta+p}^2$  is finite, the prior on  $f$  induces the prior on  $Kf$  such that  $(Kf)_i \sim N(0, \lambda_i \kappa_i^2)$ , and one can deduce from the results of Zhao (2000) and Belitser and Ghosal (2003) that

$$\sup_{\|Kf_0\|_{\beta+p} \leq R} E_0 \Pi(f : \|Kf - Kf_0\| \geq M_n n^{-\frac{(\alpha \wedge \beta) + p}{1+2\alpha+2p}} \mid Y^n) \rightarrow 0.$$

In order to apply Theorem 4.1 we need to construct the sequence of sets  $\mathcal{S}_n$  and verify condition (4.3). We use the construction as in (4.5), and we verify the remaining posterior mass condition along the lines of Lemma 4.1.

**Theorem 4.2.** *Suppose the true  $f_0$  belongs to  $S^\beta$  for  $\beta > 0$ . Then for every  $R > 0$  and  $M_n \rightarrow \infty$*

$$\sup_{\|f_0\|_\beta \leq R} E_0 \Pi(f : \|f - f_0\| \geq M_n n^{-\frac{(\alpha \wedge \beta)}{1+2\alpha+2p}} \mid Y^n) \rightarrow 0.$$

*Proof.* We first note that if  $\|f\|_\beta \leq R$ , then  $\|Kf\|_{\beta+p} \leq CR$ . Next we verify the condition of Lemma 4.1. Let

$$k_n = n^{\frac{1}{1+2\alpha+2p}}, \quad \rho_n = n^{-\frac{(\alpha \wedge \beta)}{1+2\alpha+2p}}, \quad \epsilon_n = n^{-\frac{(\alpha \wedge \beta) + p}{1+2\alpha+2p}}.$$

Note that

$$n\epsilon_n^2 = n \cdot n^{-\frac{2(\alpha \wedge \beta) + 2p}{1+2\alpha+2p}} = n^{\frac{1+2\alpha-2(\alpha \wedge \beta)}{1+2\alpha+2p}} = \epsilon_n^{-\frac{1+2\alpha-2(\alpha \wedge \beta)}{(\alpha \wedge \beta) + p}},$$

hence  $\Pi(B_n(Kf_0, \epsilon_n)) \gtrsim \exp(-C_2 n\epsilon_n^2)$  by Lemma 4.3 uniformly over a Sobolev ball of radius  $R$ ,  $S^\beta(R)$ .

Note also that

$$\rho_n^2 k_n^{1+2\alpha} = n^{-\frac{2(\alpha \wedge \beta)}{1+2\alpha+2p}} \cdot n^{\frac{1+2\alpha}{1+2\alpha+2p}} = n^{\frac{1+2\alpha-2(\alpha \wedge \beta)}{1+2\alpha+2p}} = n\epsilon_n^2,$$

and given  $c \geq 2(1+2\alpha)/\alpha$  we have  $\Pi(\mathcal{S}_n^c) \leq \exp(-(c/8)n\epsilon_n^2)$  by Lemma 4.2. Hence

$$\frac{\Pi(\mathcal{S}_n^c)}{\Pi(B_n(Kf_0, \epsilon_n))} \lesssim \exp\left(-\left(\frac{c}{8} - C_2\right)n\epsilon_n^2\right),$$

uniformly over a ball of radius  $R$ . The condition of Lemma 4.1 is verified upon choosing  $c = 8(2 + C_2) \vee 2(1 + 2\alpha)/\alpha$ .

Finally, we note that (cf. (4.6))

$$\begin{aligned} \omega(\mathcal{S}_n, f_0, \|\cdot\|, \|\cdot\|, M_n\epsilon_n) \\ &\lesssim M_n n^{\frac{p}{1+2\alpha+2p}} \cdot n^{-\frac{(\alpha \wedge \beta)+p}{1+2\alpha+2p}} + n^{-\frac{(\alpha \wedge \beta)}{1+2\alpha+2p}} + n^{-\frac{\beta}{1+2\alpha+2p}} \\ &\lesssim M_n n^{-\frac{(\alpha \wedge \beta)}{1+2\alpha+2p}}, \end{aligned}$$

which ends the proof.  $\square$

The upper bound on the posterior contraction rate in this theorem agrees with the results of Knapik et al. (2011) and Proposition 3.5 in Ray (2013). One could obtain the rate of contraction exactly as in Knapik et al. (2011), that is with scaled priors. However, this would require a refined version of Lemma 4.3, and the rate of posterior contraction for direct problem based on scaled priors. We therefore decided to set the scaling  $\tau_n \equiv 1$  and refer to the existing results in Zhao (2000) and Belitser and Ghosal (2003).

Our result on posterior contraction in the mildly ill-posed case presented in this section is not too much different from Proposition 3.5 in Ray (2013). We note three important differences: in our approach we use the existing results on posterior contraction in the direct problem, and the proofs of bounds on prior mass of the sequence  $\mathcal{S}_n$  and Kullback–Leibler type neighborhoods are elementary. Finally, our result is uniform over Sobolev balls of given radius.

**Lemma 4.2.** *Let  $\rho_n$  be an arbitrary sequence tending to 0,  $c$  be an arbitrary constant, and let the sequence  $k_n \rightarrow \infty$  satisfy  $k_n^{2\alpha} \geq 2(1 + 2\alpha)/(\alpha c \rho_n^2)$ . Then*

$$\Pi(\mathcal{S}_n^c) \leq \exp\left(-\frac{c}{8}\rho_n^2 k_n^{1+2\alpha}\right).$$

*Proof.* For  $W_1, W_2, \dots$  independent standard normal random variables

$$\Pi(\mathcal{S}_n^c) = \Pr\left(\sum_{i > k_n} \lambda_i W_i^2 > c \rho_n^2\right).$$

For some  $t > 0$

$$\begin{aligned}
& \Pr\left(\sum_{i>k_n} \lambda_i W_i^2 > c\rho_n^2\right) \\
&= \Pr\left(\exp\left(t \sum_{i>k_n} \lambda_i W_i^2\right) > \exp(tc\rho_n^2)\right) \leq \exp(-tc\rho_n^2) \mathbb{E} \exp\left(t \sum_{i>k_n} \lambda_i W_i^2\right) \\
&= \exp(-tc\rho_n^2) \prod_{i>k_n} \mathbb{E} \exp(t\lambda_i W_i^2) = \exp(-tc\rho_n^2) \prod_{i>k_n} (1 - 2t\lambda_i)^{-1/2}.
\end{aligned}$$

We first applied Markov's inequality, and later used properties of the moment generating function. Here we additionally assume that  $2t\lambda_i < 1$  for  $i > k_n$ .

We take the logarithm of the right-hand side of the previous display. Since  $\log(1 - y) \geq -y/(1 - y)$ , we have

$$\begin{aligned}
& -tc\rho_n^2 + \sum_{i>k_n} \log(1 - 2t\lambda_i)^{-1/2} \\
&= -tc\rho_n^2 - \frac{1}{2} \sum_{i>k_n} \log(1 - 2t\lambda_i) \leq -tc\rho_n^2 + \frac{1}{2} \sum_{i>k_n} \frac{2t\lambda_i}{1 - 2t\lambda_i}.
\end{aligned}$$

We continue with the latter term, noticing that  $1 - 2t\lambda_i > 1 - 2tk_n^{-1-2\alpha}$  for  $i > k_n$

$$\frac{1}{2} \sum_{i>k_n} \frac{2t\lambda_i}{1 - 2t\lambda_i} \leq \frac{t}{1 - 2tk_n^{-1-2\alpha}} \sum_{i>k_n} i^{-1-2\alpha}.$$

Since  $x^{-1-2\alpha}$  is decreasing, we have that

$$\sum_{i>k_n} i^{-1-2\alpha} \leq \int_{k_n}^{\infty} x^{-1-2\alpha} dx + k_n^{-1-2\alpha} = \frac{k_n^{-2\alpha}}{2\alpha} + k_n^{-1-2\alpha} \leq k_n^{-2\alpha} \frac{1 + 2\alpha}{2\alpha},$$

noting that  $k_n > 1$  for  $n$  large enough. Finally

$$-tc\rho_n^2 + \sum_{i>k_n} \log(1 - 2t\lambda_i)^{-1/2} \leq -tc\rho_n^2 + \frac{1 + 2\alpha}{2\alpha} \frac{t}{1 - 2tk_n^{-1-2\alpha}} k_n^{-2\alpha}.$$

Thus for  $t = k_n^{1+2\alpha}/4$

$$\Pi(\mathcal{S}_n^c) \leq \exp\left(-\frac{c}{4}\rho_n^2 k_n^{1+2\alpha} + \frac{1 + 2\alpha}{4\alpha} k_n\right) \leq \exp\left(-\frac{c}{8}\rho_n^2 k_n^{1+2\alpha}\right),$$

since  $k_n^{2\alpha} \geq 2(1 + 2\alpha)/(\alpha c\rho_n^2)$ . □

**Lemma 4.3.** *Suppose  $f_0 \in S^\beta$ . Then for every  $R > 0$  there exist positive constants  $C_1, C_2$  such that for all  $\epsilon \in (0, 1)$ ,*

$$\inf_{\|f_0\|_\beta \leq R} \Pi(B_n(Kf_0, \epsilon)) \geq C_1 \exp\left(-C_2 \epsilon^{-\frac{1+2\alpha-2(\alpha \wedge \beta)}{(\alpha \wedge \beta)+p}}\right).$$

*Proof.* This proof is adapted from Belitser and Ghosal (2003). Recall that in the white noise model the  $\ell_2$  balls and Kullback–Leibler neighborhoods are equivalent. By independence, for any  $N$ ,

$$\begin{aligned} & \Pi\left(\sum_{i=1}^{\infty}(\kappa_i f_i - \kappa_i f_{0,i})^2 \leq \epsilon^2\right) \\ & \geq \Pi\left(\sum_{i=1}^N(\kappa_i f_i - \kappa_i f_{0,i})^2 \leq \epsilon^2/2\right) \Pi\left(\sum_{i=N+1}^{\infty}(\kappa_i f_i - \kappa_i f_{0,i})^2 \leq \epsilon^2/2\right). \end{aligned} \quad (4.9)$$

Also

$$\sum_{i=N+1}^{\infty}(\kappa_i f_i - \kappa_i f_{0,i})^2 \leq 2 \sum_{i=N+1}^{\infty} \kappa_i^2 f_i^2 + 2 \sum_{i=N+1}^{\infty} \kappa_i^2 f_{0,i}^2. \quad (4.10)$$

The second sum in the display above is less than or equal to

$$2N^{-2\beta-2p} \sum_{i=N+1}^{\infty} i^{2\beta} f_{0,i}^2 \leq 2N^{-2\beta-2p} \|f_0\|_{\beta}^2 < \frac{\epsilon^2}{4},$$

whenever  $N > N_1 = (8\|f_0\|_{\beta}^2)^{1/(2\beta+2p)} \epsilon^{-1/(\beta+p)}$ .

By Chebyshev's inequality, the first sum on the right-hand side of (4.10) is less than  $\epsilon^2/4$  with probability at least

$$1 - \frac{8}{\epsilon^2} \sum_{i=N+1}^{\infty} \mathbb{E}_{\Pi}(\kappa_i^2 f_i^2) = 1 - \frac{8}{\epsilon^2} \sum_{i=N+1}^{\infty} i^{-1-2\alpha-2p} \geq 1 - \frac{4}{(\alpha+p)N^{2(\alpha+p)}\epsilon^2} > 1/2$$

if  $N > N_2 = (8/(\alpha+p))^{1/(2\alpha+2p)} \epsilon^{-1/(\alpha+p)}$ .

To bound the first term in (4.9) we apply Lemma 6.2 in Belitser and Ghosal (2003) with  $\xi_i = \kappa_i f_{0,i}$  and  $\delta^2 = \epsilon^2/2$ . Note that

$$\begin{aligned} \sum_{i=1}^N i^{1+2\alpha+2p} \xi_i^2 &= \sum_{i=1}^N i^{1+2\alpha+2p} \cdot i^{-2p} f_{0,i}^2 \\ &= \sum_{i=1}^N i^{1+2\alpha-2\beta} f_{0,i}^2 i^{2\beta} \leq N^{(1+2\alpha-2\beta) \wedge 0} \|f_0\|_{\beta}^2. \end{aligned}$$

Therefore,

$$\begin{aligned} & \Pi\left(\sum_{i=1}^N(\kappa_i f_i - \kappa_i f_{0,i})^2 \leq \epsilon^2/2\right) \\ & \geq \exp\left(-\left(1 + 2\alpha + 2p + \frac{\log 2}{2}\right)N\right) \exp\left(-N^{(1+2\alpha-2\beta) \wedge 0} \|f_0\|_{\beta}^2\right) \\ & \quad \times \Pr\left(\sum_{i=1}^N V_i^2 \leq 2\delta^2 N^{1+2\alpha+2p}\right). \end{aligned}$$

The last term, by the central limit theorem, is at least  $1/4$  if  $2\delta^2 N^{1+2\alpha+2p} > N$  and  $N$  is large, that is,  $N > N_3 = \epsilon^{-1/(\alpha+p)}$  and  $N > N_4$ , where  $N_4$  does not depend on  $f_0$ . Choosing  $N = \max\{N_1, N_2, N_3, N_4\}$  we obtain

$$\begin{aligned} \Pi(f : \|Kf - Kf_0\| \leq \epsilon) \\ \geq \frac{1}{8} \exp\left(-\left(1 + 2\alpha + 2p + \frac{\log 2}{2}\right)N\right) \exp\left(-N^{(1+2\alpha-2\beta)\wedge 0} \|f_0\|_\beta^2\right). \end{aligned}$$

Consider  $\alpha \geq \beta$ . Then  $\exp(-N) \geq \exp(-N^{(1+2\alpha-2\beta)})$  so

$$\Pi(f : \|Kf - Kf_0\| \leq \epsilon) \geq \frac{1}{8} \exp\left(-C_3 N^{(1+2\alpha-2\beta)}\right),$$

for some constant  $C_3$  that depends only on  $\alpha, \beta, p$  and  $\|f_0\|_\beta^2$ . Moreover, since  $\epsilon < 1$  and  $\alpha \geq \beta$ ,  $N$  is dominated by  $\epsilon^{-1/(\beta+p)}$  and we can write

$$\Pi(f : \|Kf - Kf_0\| \leq \epsilon) \geq \frac{1}{8} \exp\left(-C_4 \epsilon^{-\frac{1+2\alpha-2\beta}{\beta+p}}\right),$$

where  $C_4$  depends on  $f_0$  again through  $\|f_0\|_\beta^2$  only.

Now consider  $\alpha < \beta$ . Similar arguments lead to

$$\Pi(f : \|Kf - Kf_0\| \leq \epsilon) \geq \frac{1}{8} \exp\left(-C_5 \epsilon^{-\frac{1}{\alpha+p}}\right),$$

for some constant  $C_5$  that depends only on  $\alpha, \beta, p$  and  $\|f_0\|_\beta^2$ .  $\square$

#### 4.4.1.2 Severely and extremely ill-posed problems

In this section we consider the white noise setting with truncated Gaussian priors. The main purpose of this part is to show that in some classes of ill-posed problems adaptation does not need to be achieved simultaneously in both direct and indirect problems. As a matter of fact, in this part the rates in the direct problem will be much (polynomially) slower than the optimal rates. This is mostly due to the fact that we consider in here severely and extremely ill-posed problems that yield logarithmic rates of recovery. See also Knapik et al. (2013) and Agapiou et al. (2014) for examples and references.

We again consider the white noise setting, where we observe an infinite sequence  $Y^n = (Y_1, Y_2, \dots)$  as in (4.7) where  $\kappa_i \asymp \exp(-\gamma i^p)$  for some  $p \geq 1$  and  $\gamma > 0$ . Let  $Kf$  denote the sequence  $\kappa_i f_i$ , and the  $\ell_2$ -norm is denoted by  $\|\cdot\|$ . In this setting  $\mathbb{H}_1 = \mathbb{H}_2 = \ell_2$ .

We first consider estimation of  $Kf_0$  that will be later used to obtain the rate of contraction of the posterior around  $f_0$ . We put a product prior on  $f$  of the form

$$\Pi = \bigotimes_{i=1}^{k_n} N(0, \lambda_i),$$

where  $\lambda_i = i^{-\alpha} \exp(-\xi i^p)$ , for  $\alpha \geq 0$ ,  $\xi > -2\gamma$ , and some  $k_n \rightarrow \infty$ . We choose  $k_n$  solving  $1 = n\lambda_i \exp(-2\gamma i^p) = ni^{-\alpha} \exp(-(\xi + 2\gamma)i^p)$ . Using the Lambert function  $W$  one can show that

$$k_n = \left( \frac{\alpha}{p(\xi + 2\gamma)} W \left( n^{\frac{p}{\alpha}} \frac{p(\xi + 2\gamma)}{\alpha} \right) \right)^{1/p} = \left( \frac{\log n}{\xi + 2\gamma} + O(\log \log n) \right)^{1/p}, \quad (4.11)$$

see also Lemma A.4. in Knapik et al. (2013). Note that in this case we have  $\exp(k_n^p) = (nk_n^{-\alpha})^{1/(\xi+2\gamma)}$ , so we can avoid exponentiating  $k_n$ . Therefore, we do not have to specify the constant in front of the  $\log \log n$  term in the definition of  $k_n$ , and we may assume that it is of the order  $(\log n)^{1/p}$ .

Note that the hyperparameters of the prior do not depend on  $f_0$ , but only on  $K$ , which is known. For  $\mathcal{S}_n$  as in (4.5) with  $k_n$  as above and  $c = 0$ , the prior is supported on  $\mathcal{S}_n$  and the first condition of Theorem 4.1 is trivially satisfied.

**Theorem 4.3.** *Suppose the true  $f_0$  belongs to  $S^\beta$  for  $\beta > 0$ . Then for every  $R > 0$  and  $M_n \rightarrow \infty$*

$$\sup_{\|f_0\|_\beta \leq R} \mathbb{E}_0 \Pi(f : \|f - f_0\| \geq M_n (\log n)^{-\beta/p} \mid Y^n) \rightarrow 0.$$

*Proof.* Assume for brevity that we have the exact equality  $\kappa_i = \exp(-\gamma i^p)$ . Dealing with the general case is straightforward, but makes the proofs somewhat lengthier.

Since  $Y_i | f_i \sim N(\kappa_i f_i, n^{-1})$  and  $f_i \sim N(0, \lambda_i)$  for  $i \leq k_n$ , the posterior distribution (for  $Kf$ ) can be written as  $(Kf)_i | Y^n \sim N(\sqrt{nt_{i,n}} Y_i, s_{i,n})$  for  $i \leq k_n$ , where

$$s_{i,n} = \frac{\lambda_i \kappa_i^2}{1 + n\lambda_i \kappa_i^2}, \quad t_{i,n} = \frac{n\lambda_i^2 \kappa_i^4}{(1 + n\lambda_i \kappa_i^2)^2}.$$

Since the posterior is Gaussian, we have

$$\int \|Kf - Kf_0\|^2 d\Pi(Kf | Y^n) = \|\widehat{Kf} - Kf_0\|^2 + \sum_{i \leq k_n} s_{i,n}, \quad (4.12)$$

where  $\widehat{Kf}$  denotes the posterior mean and can be rewritten as:

$$\begin{aligned} \widehat{Kf} &= \left( \frac{n\lambda_i \kappa_i^2}{1 + n\lambda_i \kappa_i^2} Y_i \right)_{i=1}^{k_n} = \left( \frac{n\lambda_i \kappa_i^3 f_{0,i}}{1 + n\lambda_i \kappa_i^2} + \frac{\sqrt{n}\lambda_i \kappa_i^2 Z_i}{1 + n\lambda_i \kappa_i^2} \right)_{i=1}^{k_n} \\ &=: \mathbb{E} \widehat{Kf} + (\sqrt{t_{i,n}} Z_i)_{i=1}^{k_n}. \end{aligned}$$

By Markov's inequality the left side of (4.12) is an upper bound to  $M_n^2 \varepsilon_n^2$  times the desired posterior probability. Therefore, in order to show that  $\Pi(f : \|Kf - Kf_0\| \geq M_n \varepsilon_n | Y^n)$  goes to zero in probability, it suffices to show that the

expectation (under the true  $f_0$ ) of the right hand side of (4.12) is bounded by a multiple of  $\varepsilon_n^2$ . The last term is deterministic. As for the first term we have

$$\mathbb{E}\|\widehat{Kf} - Kf_0\|^2 = \|\mathbb{E}\widehat{Kf} - Kf_0\|^2 + \sum_{i \leq k_n} t_{i,n}.$$

We also observe

$$\|\mathbb{E}\widehat{Kf} - Kf_0\|^2 = \sum_{i \leq k_n} \frac{\kappa_i^2 f_{0,i}^2}{(1 + n\lambda_i \kappa_i^2)^2} + \sum_{i > k_n} \kappa_i^2 f_0^2.$$

We are interested in the asymptotics of the three sums

$$\sum_{i \leq k_n} \frac{\kappa_i^2 f_{0,i}^2}{(1 + n\lambda_i \kappa_i^2)^2} + \sum_{i > k_n} \kappa_i^2 f_{0,i}^2, \quad \sum_{i \leq k_n} s_{i,n}, \quad \sum_{i \leq k_n} t_{i,n}.$$

The following bounds are proven in Lemma 4.4:

$$\begin{aligned} \sum_{i \leq k_n} \frac{\kappa_i^2 f_{0,i}^2}{(1 + n\lambda_i \kappa_i^2)^2} + \sum_{i > k_n} \kappa_i^2 f_{0,i}^2 &\lesssim \|f_0\|_\beta^2 n^{-\frac{2\gamma}{\xi+2\gamma}} (\log n)^{-\frac{2\beta}{p} + \frac{2\gamma\alpha}{p(\xi+2\gamma)}}, \\ \sum_{i \leq k_n} s_{i,n} &\asymp \sum_{i \leq k_n} t_{i,n} \asymp n^{-1} (\log n)^{\frac{1}{p}}. \end{aligned} \tag{4.13}$$

Therefore, the posterior contraction rate for the direct problem is given by

$$\varepsilon_n = (\log n)^{-\frac{\beta}{p} + \frac{\gamma\alpha}{p(\xi+2\gamma)}} n^{-\frac{\gamma}{\xi+2\gamma}}.$$

By (4.6) an upper bound for the modulus of continuity is given by

$$\begin{aligned} \omega(\mathcal{S}_n, f_0, \|\cdot\|, \|\cdot\|, M_n \varepsilon_n) &\lesssim M_n \exp(\gamma k_n^p) \varepsilon_n + k_n^{-\beta} \\ &\lesssim M_n n^{\frac{\gamma}{\xi+2\gamma}} (\log n)^{-\frac{\gamma\alpha}{p(\xi+2\gamma)}} \varepsilon_n + (\log n)^{-\frac{\beta}{p}}, \\ &\lesssim M_n (\log n)^{-\frac{\beta}{p}}, \end{aligned}$$

which ends the proof.  $\square$

As already mentioned, this theorem, or rather its proof, shows that the adaptation to the optimal rate does not need to be attained simultaneously in the direct and in the inverse problem. The upper bound for the rate of contraction in the direct problem is much slower than the optimal rate of estimation of the analytically smooth parameter  $Kf_0$ , that is  $n^{-1/2}(\log n)^{1/2p}$ . This is presumably not surprising since the prior puts mass on analytic functions, whereas the true  $f_0$  belongs to the Sobolev class. There is only one choice of the parameters of the prior, namely

$\xi = 0$  and  $\alpha = \beta$  and the corresponding  $k_n$ , leading to the optimal rate also in the direct problem. This prior, however, depends on the true smoothness of  $f_0$ .

On the other hand, regardless of the choice of  $\xi$  and  $\alpha$  we achieve the optimal minimax rate of contraction  $(\log n)^{-\beta/p}$  for the inverse problem of estimating  $f_0$  (cf. Knapik et al. (2013) or Agapiou et al. (2014) and references therein). We note that other papers on Bayesian approach to severely and extremely ill-posed inverse problems do not consider truncated priors. In Knapik et al. (2013) the optimal rate is achieved for the priors with exponentially decaying or polynomially decaying variances (in the latter case the speed of decay leading to optimal rate is closely related to the regularity of the truth). Ray (2013) and Agapiou et al. (2014) obtain similar results for the priors with polynomially decaying variances. However, in the former case the rate for undersmoothing priors is worse than the rate obtained in the other papers.

We end this section with an auxiliary result used in the proof of the main result of this section.

**Lemma 4.4.** *The inequalities in (4.13) hold.*

*Proof.* Note that  $t_{i,n} \leq n^{-1}$  and  $s_{i,n} \leq n^{-1}$ . Therefore, the last two sums in (4.13) are bounded from above by  $n^{-1}k_n = n^{-1}(\log n)^{1/p}$ .

As for the first term in the first sum in (4.13) we have

$$\begin{aligned} \sum_{i \leq k_n} \frac{\kappa_i^2 f_{0,i}^2}{(1 + n\lambda_i \kappa_i^2)^2} &\leq n^{-2} \sum_{i \leq k_n} \lambda_i^{-2} \kappa_i^{-2} i^{-2\beta} i^{2\beta} f_{0,i}^2 \\ &= n^{-2} \sum_{i \leq k_n} i^{2(\alpha-\beta)} \exp(2(\xi + \gamma)i^p) i^{2\beta} f_{0,i}^2, \end{aligned}$$

and for  $k_n$  large enough all terms  $i^{2(\alpha-\beta)} \exp(2(\xi + \gamma)i^p)$  are dominated by  $k_n^{2(\alpha-\beta)} \exp(2(\xi + \gamma)k_n^p)$ , so

$$\sum_{i \leq k_n} \frac{\kappa_i^2 f_{0,i}^2}{(1 + n\lambda_i \kappa_i^2)^2} \leq n^{-2} k_n^{2(\alpha-\beta)} \exp(2(\xi + \gamma)k_n^p) \|f_0\|_\beta^2. \quad (4.14)$$

As for the second term in the first sum in (4.13) we note that

$$\sum_{i > k_n} \kappa_i^2 f_{0,i}^2 = \sum_{i > k_n} \exp(-2\gamma i^p) i^{-2\beta} i^{2\beta} f_{0,i}^2,$$

and since  $\exp(-2\gamma i^p) i^{-2\beta}$  is monotone decreasing

$$\sum_{i > k_n} \kappa_i^2 f_{0,i}^2 \leq \exp(-2\gamma k_n^p) k_n^{-2\beta} \|f_0\|_\beta^2. \quad (4.15)$$

Recall that  $\exp(k_n^p) = (nk_n^{-\alpha})^{1/(\xi+2\gamma)}$  and therefore we can rewrite the bounds in (4.14) and (4.15) as

$$n^{-2}k_n^{2(\alpha-\beta)}(nk_n^{-\alpha})^{\frac{2(\xi+\gamma)}{\xi+2\gamma}} = n^{-\frac{2\gamma}{\xi+2\gamma}}k_n^{-2\beta+\frac{2\gamma\alpha}{\xi+2\gamma}},$$

and

$$k_n^{-2\beta}(nk_n^{-\alpha})^{-\frac{2\gamma}{\xi+2\gamma}} = n^{-\frac{2\gamma}{\xi+2\gamma}}k_n^{-2\beta+\frac{2\gamma\alpha}{\xi+2\gamma}}.$$

Finally, since  $k_n$  in this case can be taken of the order  $(\log n)^{1/p}$ , we obtain the desired upper bound.  $\square$

#### 4.4.2 Regression

We now consider the inverse regression model with Gaussian residuals

$$Y_i = (Kf)(x_i) + \sigma\epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, 1) \quad (4.16)$$

where the covariate  $x_i \in \mathbb{R}$  are fixed in a covariate space  $\mathcal{X}$ . In the sequel, we take either  $\mathcal{X} = [0, 1]$  or  $\mathcal{X} = \mathbb{R}$ . In the following we consider the noise level  $\sigma > 0$  to be known although one could also think of putting a prior on it and estimate it in the direct model. In this setting, a common choice for the metric  $d$  and  $d_K$  is

$$d(f, g)^2 = n^{-1} \sum_{i=1}^n (f(x_i) - g(x_i))^2 = \|f - g\|_n^2, \quad d_K(f, g) = d(Kf, Kg).$$

For  $f \in L_2$  we denote the standard  $L_2$  norm by

$$\|f\| = \left( \int f^2 \right)^{1/2},$$

and for all  $k \in \mathbb{N}^*$ ,  $a \in \mathbb{R}^k$  we denote the usual Euclidean norm by

$$\|a\|_k = \left( \sum_{i=1}^k a_i^2 \right)^{1/2}$$

There are many known results on concentration rate of the posterior distribution for the direct model in this case, see for instance Ghosal and van der Vaart (2007) give some general conditions on the prior to achieve a certain rate. Posterior concentration rate for inverse problems has not been considered in this setting.

#### 4.4.2.1 Numerical differentiation using spline prior

In this section, we consider the inverse regression problem (4.16) with the Volterra operator defined for all measurable function  $f$  such that  $\int_0^1 f < \infty$  and  $x \in [0, 1]$  as

$$Kf(x) = \int_0^x f(t)dt. \quad (4.17)$$

This model is particularly useful for numerical differentiation for instance and has been well studied in the literature. In particular, Cavalier (2008) shows that the SVD basis for this problem is the Fourier basis and that the problem is mildly ill-posed of degree 1. We will consider a prior on  $f$  that is well suited for if the true regression function  $f_0$  belongs to the Hölder space  $\mathcal{H}(\beta, L)$  for some  $\beta > 0$ . That is  $f_0$  is  $\beta_0 = \lfloor \beta \rfloor$  times differentiable and

$$\|f_0\|_\beta = \sup_{x \neq y} \frac{|f^{(\beta_0)}(x) - f^{(\beta_0)}(y)|}{|x - y|^{\beta - \beta_0}} \leq L.$$

Since  $Kf_0$  is  $(\beta_0 + 1)$  times differentiable, it also holds that if  $f_0 \in \mathcal{H}(\beta, L)$  then  $Kf \in \mathcal{H}(\beta + 1, L)$ .

Here we construct a prior on  $f$  by considering its decomposition onto a B-splines basis. A definition of the B-spline basis can be found in De Boor (1978). For a fixed positive integer  $q > 1$  called the degree of the basis, and a given partition of  $[0, 1]$  in  $m$  subintervals of the form  $((i - 1)/m, i/m]$ , the space of splines is a collection of function  $f(0, 1] \rightarrow \mathbb{R}$  that are  $q - 2$  times differentiable and if restricted to one of the sets  $((i - 1)/m, i/m]$ , are polynomial of degree at most  $q$ . An interesting feature of the space of splines is that it forms a  $J = m + q - 1$  dimensional linear space with the so called B-spline basis denoted  $(B_{1,q}, \dots, B_{J,q})$ . Prior based on the decomposition of the function  $f$  in the B-spline basis of order  $q$  have been considered in the regression setting in Ghosal and van der Vaart (2007) and Shen and Ghosal (2014) for instance and are commonly used in practice. Here we construct a different version of the prior that will prove to be useful to derive concentration rate for the direct problem and the indirect problem. Let the prior distribution on  $f$  be defined as

$$\Pi : \begin{cases} J \sim \Pi_J \\ a_1, \dots, a_J \stackrel{iid}{\sim} \Pi_{a,J} \\ f(x) = J \sum_{j=1}^{J-1} (a_{j+1} - a_j) B_{j,q-1}(x). \end{cases} \quad (4.18)$$

Given the definition of  $B_{j,q}$  in De Boor (1978), standard computation gives

$$B'_{j,q}(x) = J (B_{j,q-1}(x) - B_{j+1,q-1}(x))$$

which in turns gives

$$Kf(x) = \sum_{j=1}^J a_j B_{j,q}(x).$$

This explains why we choose a prior as in (4.18) as it leads to the usual spline prior on  $Kf$ . Note that the condition that  $Kf(0) = 0$  can be imposed by a specific choice of nodes for the B-Splines basis (see De Boor, 1978, for more details). To compute the modulus of continuity for this model, we need to impose some conditions on the design. Let  $\Sigma_n^q$  be a matrix defined by its coefficients

$$(\Sigma_n^q)_{i,j} = \frac{1}{n} \sum_{l=1}^n B_{i,q}(x_l) B_{j,q}(x_l), \quad i, j = 1, \dots, J$$

Similarly to Ghosal and van der Vaart (2007) we ask that the design points satisfy the following conditions:

**D1** for all  $\mathbf{v}_1 \in \mathbb{R}^J$

$$J^{-1} \|\mathbf{v}_1\|_J^2 \asymp \mathbf{v}_1' \Sigma_n^q \mathbf{v}_1$$

**D2** for all  $\mathbf{v}_2 \in \mathbb{R}^{J-1}$

$$(J-1)^{-1} \|\mathbf{v}_2\|_{J-1}^2 \asymp \mathbf{v}_2' \Sigma_n^{(q-1)} \mathbf{v}_2$$

where  $a \asymp b$  means that for some constants  $c, C > 0$ ,  $ca \leq b \leq Ca$ . Condition **D1** is natural when considering B-splines priors in a regression setting, and both conditions are satisfied for a wide variety of designs. Consider for instance the uniform design  $x_i = i/n$  for  $i = 1, \dots, n$ . Then given Lemma 4.2 in Ghosal et al. (2000), we get that for  $\mathbf{v}_1 \in \mathbb{R}^J$ ,  $\mathbf{v}_2 \in \mathbb{R}^{J-1}$

$$\begin{aligned} \|\mathbf{v}_1\|_J^2 J^{-1} &\lesssim \left\| \sum_{j=1}^J \mathbf{v}_{1,j} B_{j,q} \right\|^2 \lesssim \|\mathbf{v}_1\|_J^2 J^{-1} \\ \|\mathbf{v}_2\|_{J-1}^2 (J-1)^{-1} &\lesssim \left\| \sum_{j=1}^{J-1} \mathbf{v}_{2,j} B_{j,q-1} \right\|^2 \lesssim \|\mathbf{v}_2\|_{J-1}^2 (J-1)^{-1}. \end{aligned}$$

Where the constants only depend on  $q$ . Furthermore we gave that

$$\left\| \sum_{j=1}^J \mathbf{v}_{1,j} B_{j,q} \right\|^2 = \mathbf{v}_1' \Sigma_n^q \mathbf{v}_1 + O\left(\frac{1}{n}\right),$$

where the  $O(n^{-1})$  only depends on  $q$ . We get similar results

$$\left\| \sum_{j=1}^{J-1} \mathbf{v}_{2,j} B_{j,q-1} \right\|^2 = \mathbf{v}_2' \Sigma_n^{q-1} \mathbf{v}_2 + O\left(\frac{1}{n}\right).$$

Thus **D1** and **D2** are satisfied for the uniform design for all  $J = o(n)$ .

We now go on and derive conditions on the prior such that the posterior concentrates at the minimax adaptive rate (up to a  $\log(n)$  factor). Note that here the prior distribution is neither conjugate nor depends on the SVD of the operator.

**Theorem 4.4.** *Let  $Y^n = (Y_1, \dots, Y_n)$  be a sample from (4.16) with  $\mathcal{X} = [0, 1]$  and  $\Pi$  be a prior of  $f$  as defined in (4.18). Suppose that  $\Pi_J$  is such that for some constants  $c_d, c_u > 0$  and  $t \geq 0$ , for all  $J > 1$ ,*

$$e^{-c_d j \log(j)^t} \leq \Pi_J(j \leq J \leq 2j), \quad \Pi_J(J > j) \lesssim e^{-c_u j \log(j)^t} \quad (4.19)$$

*and suppose that  $\Pi_{a,J}$  is such that for all  $a_0 \in \mathbb{R}^J$ ,  $\|a_0\|_\infty \leq H$ , there exists a constant  $c_2$  depending only on  $H$  such that*

$$\Pi_{a,J}(\|a - a_0\|_J \leq \epsilon) \geq e^{-c_2 J \log(1/\epsilon)} \quad (4.20)$$

*Define  $\Theta(\beta, L, H) = \{f \in \mathcal{H}(\beta, L), \|f\|_\infty \leq H\}$ . If the design  $(x_1, \dots, x_n)$  satisfies conditions **D1** and **D2**, then for all  $L$  and for all  $\beta \leq q$  if  $f_0 \in \mathcal{H}(\beta, L)$  there exists a constant  $C > 0$  that only depends on  $q, L, H$  and  $\Pi$  such that*

$$\sup_{\beta \leq q-1} \sup_{f_0 \in \Theta(\beta, L, H)} \mathbb{E}_0 \Pi \left( \|f - f_0\| \geq C (n)^{-\beta/(2\beta+3)} \log(n)^{3r} | Y^n \right) \rightarrow 0 \quad (4.21)$$

*with  $r = \max\{t, 1\}(\beta + 1)/(2\beta + 3)$ .*

Conditions (4.19) is similar to the one considered in Shen and Ghosal (2014) for instance, and is satisfied by the Poisson or geometric distribution for instance. Condition (4.20) is satisfied for usual choices of priors such as product of independent distribution on the  $a_j$  that admits a continuous density. Similar results hold for functions that are not uniformly bounded, with additional conditions on the tails of  $\Pi_{a,J}$ . This will only require additional computation similar to those in Shen and Ghosal (2014), and will thus not be treated here.

We first compute an upper bound for the modulus of continuity. Given conditions **D1** and **D2** we get, denoting  $\Delta(a) = (a_{j+1} - a_j)_j \in \mathbb{R}^{J-1}$

$$\begin{aligned} \|f\|_n^2 &= J^2 \Delta(a)' \Sigma_n^{q-1} \Delta(a) \\ &\lesssim J^2 \frac{1}{J-1} \|\Delta(a)\|_{J-1}^2 \\ &\lesssim J^2 \frac{1}{J-1} \|a\|_J^2 \\ &\lesssim J^2 \|Kf\|_n^2. \end{aligned}$$

To apply Theorem 4.1, we first need to derive a concentration rate for  $Kf$ . Note that in this case we simply have a standard non parametric regression model with

a spline prior. This model has been extensively studied in the literature as in Ghosal and van der Vaart (2007) or de Jonge and van Zanten (2012) and we can easily adapt their results to derive minimax adaptive concentration rates.

**Lemma 4.5.** *Let  $\Pi$  be as in Theorem 4.4. Let  $Y_n$  be sampled from model 4.16 with  $f = f_0$  and assume that  $f_0 \in \Theta(\beta, L, H)$  with  $\beta \leq q - 1$ . Then there exists a constant  $C$  that only depends on  $H, L, \Pi$ , and  $q$  such that*

$$\mathbb{E}_0 \Pi(\|Kf - Kf_0\|_n \geq Cn^{-(\beta+1)/(2\beta+3)} \log(n)^r | Y_n) \rightarrow 0$$

with  $r = \max\{t, 1\}\beta/(2\beta + 1)$ .

Similar results have been proved in Shen and Ghosal (2014), however the authors do not give a direct proof of this Theorem. Here this lemma gives us directly the posterior concentration rate for the direct problem.

*Proof.* We prove Lemma 4.5 using Theorem 4 of Ghosal and van der Vaart (2007). Let  $\beta \leq q$  and  $f_0$  be in  $\mathcal{H}(\beta, L)$  and set  $\epsilon_n = Cn^{-(\beta+1)/(2\beta+3)} \log(n)^r$  with  $r = \max\{t, 1\}\beta/(2\beta + 1)$ . Set  $J_n := J_0 n \epsilon_n^2 \log(n)^{-t}$  for a fixed constant  $J_0 > 0$  and consider the sieves  $\mathcal{S}_n$  defined by

$$\mathcal{S}_n := \{J \leq J_n, a \in \mathbb{R}^J\}$$

We first control the local entropy function  $N(\epsilon, \{J, a \in \mathcal{S}_n : \|Kf - Kf_0\| \leq \epsilon_n\}, \|\cdot\|_n)$  by using the same reasoning as in the proof of Theorem 12 of Ghosal and van der Vaart (2007) for all  $J \in \mathcal{S}_n$  we get setting

$$\log(N(\epsilon, \{J, a \in \mathcal{S}_n : \|Kf - Kf_0\| \leq \epsilon_n\}, \|\cdot\|_n)) \leq n\epsilon_n^2.$$

The prior mass of the sieve is easily controlled using the condition (4.19) as

$$\Pi(\mathcal{S}_n^c) = \Pi_J(J > J_n) \leq e^{-c_u J_n \log(J_n)^t}$$

We now need to control the prior mass of Kullback–Leiber neighbourhoods of  $Kf_0$ . Note that this condition will also be useful to apply Lemma 4.1 and thus derive the concentration rate for the direct problem. Let  $B_n(Kf_0, \epsilon)$  be defined as in (4.4)

$$B_n(Kf_0, \epsilon) = \left\{ f \in \mathcal{F} : - \int p_{Kf_0} \log \frac{p_{Kf}}{p_{Kf_0}} d\mu \leq n\epsilon^2, \right. \\ \left. \int p_{Kf_0} \left( \log \frac{p_{Kf}}{p_{Kf_0}} \right)^2 d\mu \leq n\epsilon^2, \right\},$$

Using the results of section 7.3 of Ghosal and van der Vaart (2007), setting  $\tilde{J}_n = J_n \log(n)^{-r/\beta}$  we deduce that for some constant  $c$  that only depends on  $\sigma$

$$B_n(Kf_0, \epsilon_n) \supset \{\tilde{J}_n \leq J \leq 2\tilde{J}_n, \|Kf - Kf_0\|_n^2 \leq c\epsilon_n^2\}.$$

Standard approximation results on splines gives that for all  $J$  there exists a sequence  $a_0 = (a_{0,1}, \dots, a_{0,J})$  such that

$$\|Kf_0 - \sum_{j=1}^J a_{0,j} B_{j,q}\|_n \leq J^{-\beta-1} \|Kf_0\|_\beta \leq J^{-\beta-1} L.$$

Given condition **D1** on the design, we thus have that for a constant  $c' > 0$  that only depends on  $\sigma$  and  $L$

$$B_n(Kf_0, \epsilon_n) \supset \{\tilde{J}_n \leq J \leq 2\tilde{J}_n, \|a - a_0\|_{\tilde{J}_n} \leq c' \sqrt{\tilde{J}_n \epsilon_n}\}.$$

We thus derive a lower bound on the prior mass of Kullback-Leibler neighbourhood of  $Kf_0$ .

$$\begin{aligned} \Pi(B_n(Kf_0, \epsilon_n)) &\geq \Pi\left(\tilde{J}_n \leq J \leq 2\tilde{J}_n, \|a - \omega^0\|_n \geq c' \tilde{J}_n^{1/2} \epsilon_n\right) \\ &\geq e^{-\tilde{J}_n(c_d \log(\tilde{J}_n)^t + c_2 \log(\tilde{J}_n^{-1/2} \epsilon_n^{-1}))} \end{aligned}$$

We thus have for  $C_2 > 0$ ,

$$\frac{\Pi(\mathcal{S}_n^c)}{\Pi(B_n(Kf_0, \epsilon_n))} \leq e^{-C_2 J_n \log(J_n)^t}, \quad (4.22)$$

which in turns, together with Theorem 4 of Ghosal and van der Vaart (2007) ends the proof.  $\square$

We now derive the posterior concentration rate of the posterior distribution for the inverse problem. We now get an upper bound for the modulus of continuity, for  $f \in S_n$ . Standard approximation results on splines (e.g. De Boor et al. (1978)) we have that for all  $J$  there exists  $a^0 \in \mathbb{R}^J$  such that

$$\|f_0 - \sum_{j=1}^{J-1} (a_{j+1}^0 - a_j^0)(B_{j,q-1})\|_\infty \leq (J-1)^{-\beta} \|f_0\|_\infty$$

and

$$\|Kf_0 - \sum_{j=1}^J a_j^0 B_{j,q}\|_\infty \leq J^{-\beta-1} \|Kf_0\|_\infty.$$

We thus deduce that for  $J \geq 2$ ,

$$\begin{aligned} \|f - f_0\|_n &\leq \|f - f_{a^0}\|_n + \|f_{a^0} - f_0\|_n \\ &\leq CJ^{-1} \|Kf - Kf_n\| + \|f_{a^0} - f_0\|_n \\ &\leq CJ^{-1} \|Kf - Kf_0\|_n + \|Kf_{a^0} - Kf_0\|_n + \|f_{a^0} - f_0\|_n \end{aligned}$$

We can thus deduce an upper bound for the modulus of continuity

$$\omega(S_n, f_0, \|\cdot\|_n, \|\cdot\|_n, \delta) \leq J_n \delta$$

Applying Theorem 4.1 gives

$$E_0 \Pi(\|f - f_0\|_n \geq C n^{-\beta/(2\beta+3)} \log(n)^q | Y^n) \rightarrow 0$$

for  $C > 0$  a constant that only depends on  $\|f_0\|_\infty$ ,  $q \geq 0$  and  $\Pi$ .

#### 4.4.2.2 Deconvolution using mixture priors

In this section, we consider model (4.16) where  $K$  is the convolution operator in  $\mathbb{R}$ . This model is widely used in practice, especially when considering auxiliary variables in a regression setting or for image de-blurring. For a convolution kernel  $\lambda \in L_2(\mathbb{R})$  symmetric around 0, and for all  $f \in L_2(\mathbb{R})$ , we define  $K$  as

$$Kf(x) = \lambda * f(x) = \int_{\mathbb{R}} f(u) \lambda(x - u) du, \quad \forall x \in \mathbb{R}. \quad (4.23)$$

To the authors best knowledge, theoretical properties of Bayesian nonparametric approach has not been studied for this model. In this setting we consider a mixture type prior on  $f$ , and derive an upper bound for the posterior concentration rate. Mixture priors are common in the Bayesian literature, Ghosal and van der Vaart (2001), Ghosal and van der Vaart (2007) and Shen et al. (2013) consider mixtures of Gaussian kernels, Kruijer et al. (2010) consider location scale mixture and Rousseau (2010) studied mixtures of betas. Nonetheless, since they do not fit well into the usual setting based on the SVD of the operator, mixture priors have not been considered in the literature for ill-posed inverse problems. In our case, they proved particularly well suited for the deconvolution problem. Let  $Y^n = (Y_1, \dots, Y_n)$  be sampled from model (4.16) for a true regression function  $f_0 \in L_2(\mathbb{R})$  with  $\mathcal{X} = \mathbb{R}$ , and assume that for  $c_x > 0$ , for all  $i = 1, \dots, n$ ,  $x_i \in [-c_x \log(n), c_x \log(n)]$ . This assumption is equivalent to tails conditions on the design distribution in the random design setting. Our choice of prior is well suited for  $f_0$  such that for a  $\beta > 0$ ,  $f_0$  is in the Sobolev ball  $f_0 \in S^\beta(L)$ . To avoid technicalities, we will also assume that  $f_0$  has finite support, that we may choose to be  $[0, 1]$  without loss of generality. Similar results should hold for function with support on  $\mathbb{R}$  with additional assumptions on the tails of  $f_0$  but are not treated here.

For a collection of kernels  $\Psi_v$  that depend on a the parameter  $v$ , a positive integer  $J$  and a sequence of nodes  $(z_1, \dots, z_J)$  we consider the following decomposition for the regression function  $f$  in model (4.16)

$$f(\cdot) = \sum_{j=1}^J w_j \Psi_v(\cdot - z_j),$$

where  $(w_1, \dots, w_J) \in \mathbb{R}^J$  is a sequence of weight. We choose  $\Psi_j$  proportional to a Gaussian kernel of variance  $v^2$  and the uniform sequence of nodes  $z_j = j/J$  for  $j$  such that  $j/J \in [-2c_x \log(n), 2c_x \log(n)]$

$$\Psi_{j,v}(x) = \Psi_v(x - z_j) = \frac{1}{\sqrt{2\pi}v} e^{-\frac{(x-j/J)^2}{2v^2}},$$

The choice of a Gaussian kernel is fairly natural in the nonparametric literature. In our specific case it will prove to be particularly well suited. Their main advantage here is that we can easily compute Fourier transform of  $f$  and thus use the a similar approach as in section 4.3. We consider the following prior distribution on  $f$

$$\Pi := \begin{cases} J \sim \Pi_J \\ v \sim \Pi_v \\ w_1, \dots, w_J | J \sim \otimes_{j=1}^J N(0, 1) \end{cases} \quad (4.24)$$

We use a specific Gaussian prior for the weight  $(w_1, \dots, w_J)$  in order to use the results on Reproducing Kernel Hilbert Spaces following de Jonge and van Zanten (2010) to derive concentration rate for the direct problem. However our intuition is that this results should holds for a more general classes of prior but the computations would be more involved.

Following Fan (1991), we define the degree of ill-posedness of the problem through the Fourier transform of the convolution kernel. For  $p > 0$ , we say that the problem is mildly ill posed of degree  $p$  if there exists some constants  $c, C > 0$  such that for  $\hat{\lambda}$  the Fourier transform of  $\lambda$

$$\hat{\lambda}(t) = \int \lambda(u) e^{itu} du,$$

we have for  $|t|$  sufficiently large

$$c|t|^{-p} \leq |\hat{\lambda}(t)| \leq C|t|^{-p}, p \in \mathbb{N}^* \quad (4.25)$$

For all  $f_0 \in S^\beta(L)$ , we have that  $Kf_0 \in S^{\beta+p}(L')$  for  $L' = LC$ . Under these conditions, the following Theorem gives an upper bound on the posterior concentration rate.

**Theorem 4.5.** *Let  $Y^n = (Y_1, \dots, Y_n)$  be sampled from (4.16) with  $\mathcal{X} = \mathbb{R}$  and assume that the design points  $(x_i)$  are such that  $(x_i) \in [-c \log(n), c \log(n)]^n$ . Let  $f_0$  be such that for  $\beta \in \mathbb{N}^*$  and  $M > 0$ ,  $f_0 \in S^\beta(L)$  with support on  $[0, 1]$  and  $\|f_0\|_\infty \leq M$ . Consider  $K$  to be as in (4.23) with  $\lambda$  satisfying (4.25). Let  $\Pi$  be a prior distributions defined as in (4.24) with*

$$\Pi_J(J = j) \asymp j^{-s} \quad (4.26a)$$

$$v^{-q} e^{-\frac{c_d}{v} \log(1/v)^r} \lesssim \Pi_v(v) \lesssim v^{-q} e^{-\frac{c_u}{v} \log(1/v)^r}. \quad (4.26b)$$

Then there exists a constant  $C$  and  $r$  that only depends on  $\Pi$ ,  $L$ ,  $K$  and  $M$  such that

$$E_0^n \Pi(\|f - f_0\| \geq Cn^{-\beta/(2\beta+2p+1)} \log(n)^r |Y^n) \rightarrow 0,$$

as  $n$  goes to  $\infty$ .

Note that here the prior does not depend on the regularity  $\beta$  of  $f_0$ , we have the adaptive minimax concentration rates for this problem. Note also that the prior does not depend on the degree of ill-posedness either. It is thus well suited for a wide variety of convolution kernels. In particular this can be useful when the operator is only partially known, as in this case the regularity of the prior may not be accessible. However, this case is beyond the scope of this article. We prove Theorem 4.5 by applying Theorem 4.1 together with Lemma 4.1. A first difficulty is to explicit the set  $\mathcal{S}_n$  on which we can control the modulus of continuity. A second problem is to derive the posterior concentration rate for the direct problem, given that here  $Kf$  is supported on the real line. de Jonge and van Zanten (2010) derived the posterior concentration rate for Hölder smooth function with bounded support. However, their results directly extend to the case of convolution of Hölder functions with bounded support.

*Proof.* We first specify the set  $\mathcal{S}_n$  for which we can control the modulus of continuity. Denoting  $\hat{f}$  the Fourier transform of  $f$ , for any sequence  $a_n$  going to infinity and  $I_n = [-a_n, a_n]$  we define for  $a > 0$

$$\mathcal{S}_n = \left\{ f, \int_{I_n} |\hat{f}(t)|^2 dt \geq a \int_{I_n^c} |\hat{f}(t)|^2 dt \right\}. \quad (4.27)$$

We control the modulus of continuity  $\omega(\mathcal{S}_n, f_0, \|\cdot\|, \|\cdot\|, \delta)$  in a similar way as in Section 4.3. First consider  $f \in \mathcal{S}_n$ , we have denoting  $\hat{f}_n(\cdot) = \hat{f}(\cdot)\mathbb{I}_{I_n}(\cdot)$

$$\begin{aligned} \|f\|^2 &= \|\hat{f}\|^2 \\ &\leq (1+a)\|\hat{f}_n\|^2 \\ &\lesssim a_n^{2p} \int_{I_n} |\hat{f}|^2 |\hat{\lambda}|^2 \lesssim a_n^{2p} \|Kf\|^2 \end{aligned}$$

Note that for  $f_0 \in S^\beta(L)$  we have for  $f_{0,n}(x) = \int \hat{f}_{0,n}(t) e^{-itx} dt$

$$\|f_0 - f_{0,n}\| \leq 2a_n^{-\beta} L, \|Kf_0 - Kf_{0,n}\| \leq 2a_n^{-(\beta+p)},$$

which in turns gives

$$\omega(\mathcal{S}_n, f_0, \|\cdot\|, \|\cdot\|, \delta) \lesssim a_n^p \delta + a_n^{-\beta}. \quad (4.28)$$

We now control the prior mass of  $\mathcal{S}_n^c$  in order to apply Lemma 1. Denote by  $l_n = \lfloor a_n/(2\Pi J) \rfloor$ ,  $L_n = \lceil a_n/(2\Pi J) \rceil$ , we have

$$\begin{aligned}
\int_{I_n} |\hat{f}(t)|^2 dt &\geq 2\pi J \int_{-L_n}^{l_n} e^{-4\pi^2 t^2 v^2} \left| \sum_{j=1}^J w_j e^{2\pi j t} \right| dt \\
&= 2\pi J \sum_{l=-L_n}^{l_n} \int_l^{l+1} e^{-4\pi^2 t^2 v^2} \left| \sum_{j=1}^J w_j e^{2\pi j t} \right| dt \\
&= 2\pi J \int_0^1 \left| \sum_{j=1}^J w_j e^{2\pi j t} \right| \sum_{l=-L_n}^{l_n} e^{-4\pi^2 (t+l)^2 v^2} dt \\
&\geq 2\pi J \sum_{l=-L_n}^{l_n} e^{-4\pi^2 (1+l)^2 v^2} \int_0^1 \left| \sum_{j=1}^J w_j e^{2\pi j t} \right| dt
\end{aligned}$$

and similarly we get

$$\begin{aligned}
\int_{I_n^c} |\hat{f}(t)|^2 dt &\leq 2\pi J \int_0^1 \left| \sum_{j=1}^J w_j e^{2\pi j t} \right| \sum_{l=-\infty}^{-L_n} e^{-4\pi^2 (t+l)^2 v^2} + \sum_{l=l_n}^{\infty} e^{-4\pi^2 (t+l)^2 v^2} dt \\
&\leq 2\pi J \left( \sum_{l=-\infty}^{-L_n} e^{-4\pi^2 l^2 v^2} + \sum_{l=l_n}^{\infty} e^{-4\pi^2 l^2 v^2} \right) \int_0^1 \left| \sum_{j=1}^J w_j e^{2\pi j t} \right| dt.
\end{aligned}$$

We thus deduce that for an absolute constant  $C, C' > 0$

$$\Pi(\mathcal{S}_n^c) \leq \Pi(v \leq J/a_n) \lesssim e^{-C' a_n \log(a_n)}$$

We now adapt the results of de Jonge and van Zanten (2010) to our setting in order to get the control of the posterior mass of the Kullback-Leibler neighbourhoods of  $Kf_0$  and the posterior concentration rate for the direct problem. Following their notations we have that  $K\Psi_v \in \mathcal{P}_\infty$ , and thus the small ball probability  $\Pi(\|f\|_\infty \leq \epsilon)$  can be controlled by their Lemma 3.3. We extend their Lemma 3.5 to our setting. Note that with Lemma 9 of Scricciolo (2014), Lemma 3.4 of de Jonge and van Zanten (2010) holds for the same  $T_{\alpha,v}$  with  $\alpha = \beta + p$ . Choosing  $h$  to be as in the proof of Lemma 3.5 of de Jonge and van Zanten (2010) and denoting  $\omega_0 = f_0 \star \lambda$ , we have

$$h(x) = \sum_{j/J \in [-2c_x \log(n), 2c_x \log(n)]} T_{\alpha,v}(\omega_0) \frac{1}{Jv} \Psi\left(\frac{x - j/J}{v}\right),$$

and thus deduce

$$\|h\|_{H^{J,v}}^2 \leq \|T_{\alpha,v}(\omega_0)\|^2 2c_x \log(n).$$

Using their decomposition (3.8), we control  $|h(x) - \Psi_v \star T_{\alpha,v}(\omega_0)(x)|$  along the same lines as in their computations page 3312. We have

$$\begin{aligned} |h(x) - \Psi_v \star T_{\alpha,v}(\omega_0)(x)| &\leq \left| h(x) - \int_{-2c_x \log(n)}^{2c_x \log(n)} T_{\alpha,v}(\omega_0)(y) \Psi_v(x-y) dy \right| \\ &\quad + \left| \int_{-\infty}^{-2c_x \log(n)} T_{\alpha,v}(\omega_0)(y) \Psi_v(x-y) dy \right| \\ &\quad + \left| \int_{2c_x \log(n)}^{\infty} T_{\alpha,v}(\omega_0)(y) \Psi_v(x-y) dy \right| \quad (4.29) \end{aligned}$$

The first display of (4.29) can be controled as in the proof of Lemma 3.5 of de Jonge and van Zanten (2010). For the last two displays, we have

$$\begin{aligned} \left| \int_{-\infty}^{-2c_x \log(n)} T_{\alpha,v}(\omega_0)(y) \Psi_v(x-y) dy \right| + \left| \int_{2c_x \log(n)}^{\infty} T_{\alpha,v}(\omega_0)(y) \Psi_v(x-y) dy \right| \\ \lesssim \|T_{\alpha,v}(\omega_0)\|_{\infty} e^{-\frac{c_x^2 \log(n)^2}{2v^2}} v^{-1}. \end{aligned}$$

Following the same proof of Theorem 2.2 of de Jonge and van Zanten (2010), we get

$$\mathbb{E}_0 \Pi(\|Kf - Kf_0\| \geq Cn^{-(\beta+p)/(2\beta+2p+1)} \log(n)^{r_0} |Y^n) \rightarrow 0$$

and similarly to their equation (2.5) we get, with  $\epsilon_n = n^{-(\beta+p)/(2\beta+2p+1)} \log(n)^{r_0}$

$$\Pi(\|Kf - Kf_0\| \leq \epsilon_n) \geq e^{-n\epsilon_n^2}.$$

Choosing  $a_n = n\epsilon_n^2$ , together with Lemma 4.1 and Theorem 4.1, this gives us the desired results.  $\square$

## 4.5 Discussion

In this paper we propose a new approach to the problem of deriving posterior concentration rates for linear ill-posed inverse problems. More precisely, we put a prior on the parameter of interest  $f$  that naturally imposes the prior on  $Kf$ , leading to a certain rate of contraction in the direct problem. Next, we consider a sequence of sets on which the operator  $K$  possesses a continuous inverse. Then, we impose additional conditions on the prior (or the posterior itself) under which the posterior concentrates at a certain rate in the inverse problem setting.

This is a great advantage of the Bayesian approach in this setting as when the posterior distribution is known to concentrate at a given rate in the direct

problem, one only has to consider subset of high prior mass for which the norm of the inverse of the operator may be handled. Our result seems to show that the main difficulty when considering linear inverse problems is to control the change of norms from  $d_K$  to  $d$ , which is dealt here by considering the modulus of continuity as introduced in Donoho and Liu (1991) and Hoffmann et al. (2013). It is also to be noted that contrariwise to existing methods, we do not require a Hilbertian structure for the parameter space, see for instance the example treated in Section 4.4.2.1. This could be particularly useful when considering nonlinear operators, and is of potential interest when considering the case of partially known operators.

We recovered (a subset of) the existing results from Knapik et al. (2011), Knapik et al. (2013), Agapiou et al. (2013), Agapiou et al. (2014), and Ray (2013). Our approach should be viewed as a generalization of the ideas presented in the latter paper. Furthermore, we were able to derive posterior concentration rates for prior distributions that were not covered by the existing theory. In this sense, the approach proposed in this paper is more general, and we believe more natural, than the existing ones.

## Bibliography

- Agapiou, S., Larsson, S., and Stuart, A. M. (2013). Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Process. Appl.*, 123(10):3828–3860.
- Agapiou, S., Stuart, A. M., and Zhang, Y.-X. (2014). Bayesian posterior contraction rates for linear severely ill-posed inverse problems. *J. Inverse Ill-Posed Probl.*, 22(3):297–321.
- Belitser, E. and Ghosal, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.*, 31(2):536–559.
- Castillo, I. (2013). On bayesian supremum norm contraction rates. *arXiv preprint arXiv:1304.1761*.
- Cavalier, L. (2008). Nonparametric statistical inverse problems. *Inverse Problems*, 24(3):034004, 19.
- De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag New York.
- De Boor, C., De Boor, C., De Boor, C., and De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag New York.
- de Jonge, R. and van Zanten, J. H. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *Ann. Statist.*, 38(6):3300–3320.
- de Jonge, R. and van Zanten, J. H. (2012). Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electron. J. Stat.*, 6:1984–2001.

- Donoho, D. L. and Liu, R. C. (1991). Geometrizing rates of convergence, ii. *The Annals of Statistics*, 19(2):633–667.
- Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of inverse problems*, volume 375. Springer.
- Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pages 1257–1272.
- Florens, J.-P. and Simoni, A. (2012). Regularized posteriors in linear ill-posed inverse problems. *Scandinavian Journal of Statistics*, 39(2):214–235.
- Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531.
- Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.*, 35(1):192–223.
- Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5):1233–1263.
- Hoffmann, M., Rousseau, J., and Schmidt-Hieber, J. (2013). On adaptive posterior concentration rates. *arXiv preprint arXiv:1305.5270*.
- Knapik, B. T., Szabó, B. T., van der Vaart, A. W., and van Zanten, J. H. (2012). Bayes procedures for adaptive inference in nonparametric inverse problems. *Preprint*. (arXiv:1209.3628).
- Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.*, 39(5):2626–2657.
- Knapik, B. T., van der Vaart, A. W., and van Zanten, J. H. (2013). Bayesian recovery of the initial condition for the heat equation. *Comm. Statist. Theory Methods*, 42.
- Kruijer, W., Rousseau, J., and van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.*, 4:1225–1257.
- Ray, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electronic Journal of Statistics*, 7:2516–2549.
- Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.*, 38(1):146–180.
- Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710.
- Salomond, J.-B. (2013). Concentration rate and consistency of the posterior under monotonicity constraints. *arXiv preprint arXiv:1301.1898*.
- Scricciolo, C. (2014). Adaptive bayesian density estimation in  $l^p$ -metrics with pitman-yor or normalized inverse-gaussian process kernel mixtures. *Bayesian Analysis*, 9(2):475–520.

- Shen, W. and Ghosal, S. (2014). Adaptive Bayesian procedures using random series prior. *ArXiv e-prints*.
- Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive bayesian multivariate density estimation with dirichlet mixtures. *Biometrika*, 100(3):623–640.
- Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714.
- Zhao, L. H. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.*, 28(2):532–552.