Contributions to the theoretical analysis of Statistical learning and Uncertainty Quantification methods

Thibault Randrianarisoa

Contributions to the theoretical analysis of Statistical learning and Uncertainty Quantification methods

par

Thibault Randrianarisoa

Thèse soutenue le **28 septembre, 2022**

En vue de l'obtention du titre de
Docteur de Sorbonne Université
Discipline: ***Mathématiques***
Spécialité ***Statistique***
École Doctorale Sciences Mathématiques de Paris Centre
Laboratoire de Probabilités, Statistique et Modélisation - LPSM

Devant un jury composé de:

Rapporteurs:	Edward I. George	University of Pennsylvania
	Gilles Blanchard	Université Paris Saclay
Directeur de thèse:	Ismaël Castillo	Sorbonne Université
Examinateurs:	Veronicka Ročková	University of Chicago
	Johannes Schmidt-Hieber	University of Twente
	Botond Szabó	Bocconi University
	Aad van der Vaart	Delft University of Technology
	Olivier Wintenberger	Sorbonne Université

SORBONNE UNIVERSITÉ

LPSM
Laboratoire de probabilités statistique & modélisation

Remerciements

Ce manuscrit n'a pu voir le jour que grâce à l'intervention et au soutien remarquable dont j'ai pu faire preuve. Je souhaite donc saluer ceux que j'ai eu la chance de côtoyer ces trois dernières années..

En premier lieu, je tiens à te remercier, Ismaël. Je suis heureux que tu m'aies fait confiance pour travailler sur les projets de cette thèse. Je m'y suis d'autant plus intéressé que tu m'as appris à prendre du recul par rapport aux mathématiques pour mieux appréhender leur subtilité et leur beauté. Tu as su m'accorder une certaine liberté de recherche, mais également te montrer disponible lorsque j'avais besoin de tes conseils. Tes réponses à mes questions et tes remarques, toujours adéquates, m'ont permis de surmonter de nombreuses difficultés. Finalement, c'est grâce à ta bienveillance que j'ai pu m'épanouir tout au long de cette aventure, si intimidante au début. Je suis profondément reconnaissant d'avoir été encadré par un chercheur aussi passionné et ces quelques mots ne sauraient exprimer toute ma gratitude.

La valeur de cette thèse tient également beaucoup au travail d'Ed et Gilles, que je remercie fortement d'avoir accepté d'en être les rapporteurs. *I am deeply grateful and honoured that you accepted to report my thesis, Ed !* Je les remercie d'avoir pris du temps pour lire ce texte en détail et m'en faire un retour minutieux. Merci également aux autres membres du jury Aad, Johannes, Olivier, Veronicka ainsi que Botond avec qui je suis impatient de collaborer pour la suite de mon parcours.

Merci à ceux avec qui j'ai eu l'occasion de discuter et de travailler sur divers projets: Agnès, François, Neil et Richard. Ce fut un plaisir d'échanger avec vous ! Tout au long de ma thèse, j'ai pu rencontrer des personnes enthousiasmées par la recherche et toujours ravies de discuter et d'expliquer leurs travaux. Merci à elles, pour leur passion et pour les discussions enrichissantes qui en ont suivi. Je pense ici entre autres à mon ami Dorian, que j'ai pu recroiser malgré son exil dans le territoire des bandits du Nord.

À mes amis qui m'accompagnent depuis le début de ma thèse, je dis "*gracias*" Gloria, et "*tak*", Nicklas! *You played a significant role in the office's joyful atmosphere, and I cherished our moment of laughter !* Je remercie tous ceux qui ont partagé le bureau 206 avec nous, en particulier Aude, Cyril, Grâce et Qiming. J'adresse également mes remerciements aux nombreux doctorants du laboratoire : Adeline, Alexis, Alice, Antonio, Ariane, Camila, Franceso, Iqraa, Joseph, Ludovic, Miguel, Nicolas, Pierre, Sébastien... et aux permanents qui m'ont accompagné à un moment ou à un autre : Anna, Antoine, Charlotte, Eddie, Étienne, Gérard, Maud, Olivier, Tabea. Je remercie l'équipe organisatrice du GTT: David, Lucas, Émilien et Yoan, avec qui nous avons passé une année bien sympathique. Merci, Hugues, pour ton aide et tes mails décalés, mais toujours rafraîchissants. Merci au secrétariat du laboratoire pour votre aide et votre amabilité: Corinne, Élise, Fatime, Florence, Louise, Nathalie et Valérie. De manière générale, je voudrais remercier l'ensemble des collègues du LPSM qui en ont fait un cadre de travail agréable.

Je considère que mes heures de loisir ont été tout aussi importantes que mes heures de travail dans l'aboutissement de ce manuscrit. Dans ces moments de détente, j'ai pu compter sur mes amis. Antoine D., j'ai entre $7,50$ et $8,6$ raisons de te remercier, mais je ne saurai toutes les énumérer. Antoine V., nos énièmes reprises de course ont été un vrai bol d'air frais (beaucoup plus que nos parties de FIFA). Guillaume, mon ami de longue date et véritable compagnon de thèse : on pourra dire qu'on aura parcouru beaucoup de chemin depuis ce jour où l'on a déchiffré ton grimoire d'allemand !

Enfin, je remercie fortement ma famille pour leur soutien inconditionnel. J'ai une pensée toute particulière pour vous, papa et maman ; vous qui m'avez toujours encouragé dans chacune de mes entreprises et vous êtes investi tout autant que moi dans mes études. Je n'ai eu de cesse de m'efforcer de vous rendre fiers et je mesure la chance que j'ai d'avoir des parents aussi aimants. Flavien et Manon, nos moments de rire et de partage ont été d'un réconfort inestimable dans mes moments de doutes et je chéris chaque jour notre complicité. Je vous dédie ce travail en partie. Je pense également à mes grands-parents qui auront su me procurer tant de souvenirs d'enfance et auront été parmi mes supporters de la première heure. Pour finir, je te remercie aussi, Claire, pour ton soutien dans cette dernière ligne droite et pour beaucoup d'autres aventures à venir.

Abstract

Modern data analysis provides scientists with statistical and machine learning algorithms with impressive performance. In front of their extensive use to tackle problems of constantly growing complexity, there is a real need to understand the conditions under which algorithms are successful or bound to fail. An additional objective is to gain insights into the design of new algorithmic methods able to tackle more innovative and challenging tasks. A natural framework for developing a mathematical theory of these methods is *nonparametric inference*. This area of Statistics is concerned with inferences of unknown quantities of interest under minimal assumptions, involving an infinite-dimensional statistical modeling of a parameter on the data-generating mechanism. In this thesis, we consider both problems of function estimation and uncertainty quantification.

The first class of algorithms we deal with are Bayesian tree-based methods. They are based on a 'divide-and-conquer' principle, partitioning a sample space to estimate the parameter locally. In regression, these methods include BCART and the renowned BART, the later being an ensemble of trees or a forest. In density estimation, the famous Pólya Tree prior exemplifies these methods and is the building block of a myriad of related constructions. We propose a new extension, DPA, that is a 'forest of PTs' and is shown to attain minimax contraction rates adaptively in Hellinger distance for arbitrary Hölder regularities. Adaptive rates in the stronger supremum norm are also obtained for the flexible Optional Pólya Tree (OPT) prior, a BCART-type prior, for regularities smaller than one.

Gaussian processes are another popular class of priors studied in Bayesian nonparametrics and Machine Learning. Motivated by the ever-growing size of datasets, we propose a new horseshoe Gaussian process with the aim to adapt to leverage a data structure of smaller dimension. First, we derive minimax optimal contraction rates for its tempered posterior. Secondly, deep Gaussian processes are Bayesian counterparts to the famous deep neural networks. We prove that, as a building block in such a deep framework, it also gives optimal adaptive rates under compositional structure assumptions on the parameter.

As for uncertainty quantification (UQ), Bayesian methods are often praised for the principled solution they offer with the definition of credible sets. We prove that OPT credible sets are confidence sets with good coverage and size (in supremum norm) under qualitative self-similarity conditions. Moreover, we conduct a theoretical study of UQ in Wasserstein distances W_p, uncovering a new phenomenon. In dimensions smaller than 4, it is possible to construct confidence sets whose W_p-radii, $p \leq 2$, adapt to any regularities (with no qualitative assumptions). This starkly contrasts the usual L_p theory, where concessions always have to be made.

Keywords: Bayesian nonparametrics, Tree-based methods, Uncertainty Quantification, Wasserstein distance, Gaussian process

Résumé

L'analyse moderne des données fournit aux scientifiques des algorithmes statistiques et d'apprentissage automatique aux performances impressionnantes. Face à leur utilisation intensive pour traiter des problèmes dont la complexité ne cesse de croître, il existe un réel besoin de comprendre les conditions dans lesquelles ceux-ci fonctionnent ou sont voués à l'échec. Ainsi, un cadre naturel pour développer une théorie mathématique de ces méthodes est celui de l'inférence non-paramétrique. Ce domaine de la statistique s'intéresse à l'inférence de quantités inconnues sous des hypothèses minimales avec la modélisation statistique en dimension infinie d'une quantité paramétrant la loi des données. Dans cette thèse, nous étudions les problèmes d'estimation de fonctions et de quantification de l'incertitude.

La première classe d'algorithmes que nous considérons est celle des méthodes bayésiennes basée sur des structures d'arbres. Elles reposent sur le principe de 'diviser pour mieux régner', en partitionnant l'espace des données pour estimer le paramètre localement. En régression, ces méthodes incluent BCART et BART, cette dernière étant un ensemble d'arbres ou "forêt". En estimation de densité, les arbres de Pólya sont un exemple de telles lois a priori et constituent la base d'une myriade de constructions connexes. Nous proposons une nouvelle extension, DPA, qui est une "forêt de Pólya" et permet d'atteindre des vitesses de contraction minimax, de manière adaptative, en distance de Hellinger pour des régularités de Hölder arbitraires. Des vitesses adaptatives dans la norme infinie sont également obtenues pour la loi a priori des arbres de Pólya optionnel (OPT), similaire à BCART en régression, pour des fonctions de régularité Lipschitz.

Les processus gaussiens (GP) sont une autre classe populaire de lois étudiées en statistique bayésienne nonparamétrique et en apprentissage automatique. Motivés par la taille toujours croissante des bases de données, nous proposons un nouveau processus gaussien 'horseshoe' avec une couche de sélection de variables 'soft' pour pouvoir tirer parti d'une dimension des données plus petite que celle de l'espace ambiant. Nous dérivons des vitesses de contraction optimales pour les loi a posteriori tempérées. Les processus gaussiens profonds sont les homologues bayésiens des célèbres réseaux neuronaux profonds. Nous prouvons que, en tant qu'élément de base dans une telle construction, les GP 'horseshoe' donnent également des vitesses adaptatives sous des hypothèses de structure de composition du paramètre.

En ce qui concerne la quantification de l'incertitude (UQ), les méthodes bayésiennes sont souvent louées pour la solution qu'elles fournissent avec la définition des ensembles de crédibilité. Nous prouvons que ces ensembles construits sous OPT sont des ensembles de confiance avec un niveau de confiance exact et une taille optimale (ou quasi-optimale) en norme infinie sous des conditions qualitatives d'auto-similarité. De plus, nous menons une étude théorique de l'UQ pour les distances de Wasserstein W_p et mettons en lumière un nouveau phénomène. En dimensions inférieures à 4, il est toujours possible de construire des ensembles de confiance dont les rayons en distance W_p, $p \leq 2$, s'adaptent à n'importe quelles régularités (sans

hypothèses qualitatives). Cela contraste fortement avec la théorie habituelle en norme L_p, où des concessions doivent toujours être faites.

Keywords: Bayésien non-paramétrique, méthodes par arbres, quantification de l'incertitude, distances de Wasserstein, processus gaussiens

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Statistical framework

1.1.1 Nonparametric Inference

Suppose one observes a random variable $\mathbf{X}^{(n)}$, for $n \geq 1$ an integer, from a measurable space $(\mathfrak{X}, \mathfrak{A})$, with $\mathfrak{A}$ a σ-field over $\mathfrak{X}$. The law of this random variable is assumed to belong to a statistical model

$$\left\{P_f^{(n)}, f \in \mathcal{F}\right\}, \tag{1.1}$$

where the P_f's are probability measures on $\mathfrak{A}$. If f_0 denotes the true unknown value which has generated the data, the goal is to make inference (i.e., recover or at least "approximate", in a sense to be made precise below) on f_0 from $\mathbf{X}^{(n)}$, where n quantifies the amount of information available (we can think of it as a sample size, see below).

As opposed to parametric statistics, one works here with infinite-dimensional parameter spaces $\mathcal{F}$, i.e. functional spaces. Indeed, one makes the more practical and flexible assumption that the parameter f cannot be described by a finite-dimensional vector, aiming to capture finer aspects of the data generating distribution. Two examples of such settings which will be prominent in this thesis are the following.

Example 1 (Density estimation)**.** *The parameter space consists of a subset of the set of all probability densities on some sample space $(\mathcal{X}, \mathcal{A})$ and the observations can be written as $\mathbf{X}^{(n)} = (X_1, \ldots, X_n)$, with X_i independent identically distributed (henceforth i.i.d.) as P_{f_0}, where P_{f_0} denotes the absolutely continuous distribution with probability density f_0 with respect to a σ-finite measure λ on $\mathcal{A}$. The model consists of the product measures $P_f^{(n)} = P_f^{\otimes n}$ on the measurable space $(\mathfrak{X}, \mathfrak{A}) = (\mathcal{X}^{\otimes n}, \mathcal{A}^{\otimes n})$.*

Example 2 (Nonparametric Gaussian regression with random design)**.** *One observes n independent and identically distributed pairs of random variables $\mathbf{X}^{(n)} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ where $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ and*

$$Y_i = f_0(X_i) + \varepsilon_i, \ i = 1, \ldots, n.$$

There, $\mathcal{X}$ is a (possibly multidimensional) covariate space, the variables X_i are distributed given a probability measure μ, the ε_i are independent centered Gaussian random variables, independent of the design vector $(X_1, \ldots, X_n)$. The parameter space is then a subset of the

set of regression functions $f : \mathcal{X} \to \mathbb{R}$ and $P_f^{(n)}$ denotes the distribution of the n i.i.d. tuples of variables (X_i, Y_i) under the regression function f.

In the two examples above, $\mathcal{F}$ is typically a subset of a functional space, defined via a qualitative restriction. For instance, we may intersect the set of probability densities or the set of regression functions on a bounded set such as the cube $[-1,1]^d$ with the Hölder class

$$\Sigma(\beta, L) = \left\{ f : \|f\|_{\beta_i,\infty} \leq L \right\}, \tag{1.2}$$

where $L \geq 0$ and, for $\beta \in \mathbb{R}_+^{*d}$,

$$\|f\|_{\beta_i,\infty} := 2r \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|<\lfloor\beta_i\rfloor} \|\partial^{\boldsymbol{\alpha}} f\|_\infty + 2^{\beta_i - \lfloor\beta_i\rfloor} \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor\beta_i\rfloor} \sup_{\boldsymbol{x},\boldsymbol{y}\in[-1;1]^r,\ \boldsymbol{x}\neq\boldsymbol{y}} \frac{|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{y})|}{|\boldsymbol{x} - \boldsymbol{y}|_\infty^{\beta_i - \lfloor\beta_i\rfloor}}, \tag{1.3}$$

with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r) \in \mathbb{N}^r$, $|\boldsymbol{\alpha}| := |\boldsymbol{\alpha}|_1$ and $\partial^{\boldsymbol{\alpha}} = \partial^{\alpha_1} \ldots \partial^{\alpha_r}$ denotes the α-partial derivative. For increasing regularities β and decreasing radii L, $\Sigma(\beta, L)$ contains functions that are smoother and we will see the consequences in Sections 1.1.2 and 1.1.5.

In this thesis, one focuses on two relevant problems in the recovering of f_0:

- the construction of an estimator $\hat{f}_n : \mathfrak{X} \to \mathcal{F}$ such that $\hat{f}_n\left(\mathbf{X}^{(n)}\right)$ is close to f_0 when $\mathbf{X}^{(n)} \sim P_{f_0}^{(n)}$, where the notion of closeness is made precise in the next subsection;
- the construction of confidence sets $C(\mathbf{X}^{(n)})$, containing f_0 with high probability under the law of $\mathbf{X}^{(n)} \sim P_{f_0}^{(n)}$, with C being as small as possible (see Section 1.3.2 for the explicit requirements).

While building an estimate of the parameter is an important step in inference, from a practical point-of-view, estimates should preferably always be associated to an evaluation of their uncertainty. A natural way to quantify this is to account for the variability of the estimate, with regards to the randomness of the observations. This brings us to the construction of sets of 'plausible' parameter values, and called confidence sets. These contain the true parameter value with high probability, and thus inform statisticians on the reliability of their estimates.

1.1.2 Convergence rates

In order to evaluate the performance of statistical procedures applied to the above problems, we use a generalized metric d (e.g., a distance, a squared distance, a divergence...) on $\mathcal{F}$. In estimation, we would like to measure closeness between our estimator $\hat{f}_n\left(\mathbf{X}^{(n)}\right)$ and f_0, and to discard trivial estimators with excellent properties for a handful of parameters but poor performance globally, we define the uniform measure

$$r_n := \sup_{f\in\mathcal{F}} \mathbb{E}_f^{(n)}\ d\left(\hat{f}, f\right)$$

that is the *rate of convergence* of the estimator. Then, the question arises as to whether it is possible to improve upon this rate and to the definition of the *minimax rate* over the class $\mathcal{F}$

$$r_n^* := \inf_{\hat{f}} \sup_{f\in\mathcal{F}} \mathbb{E}_f^{(n)}\ d\left(\hat{f}, f\right),$$

where the infimum is over the set of estimators $\hat{f} : \mathfrak{X} \to \mathcal{F}$ and the expectation is under the law of $\mathbf{X}^{(n)} \sim P_f^{(n)}$. We note in passing that one may also define a notion of *minimax convergence rates in probability*, which we use and define in Chapter 4. We should resort to estimators converging 'as fast as possible', i.e., whose rate of convergence are of the same order as r_n^*, or the closest possible.

As more data information and information should improve their performance, it is enlightening and useful for the understanding of statistical procedures to analyze their behaviour in the limit $n \to \infty$. Such an asymptotic analysis is often easier to conduct and this can help to classify statistical algorithm given their asymptotic performance. Indeed, it may be possible that no best algorithm, in a uniform sense, exists. Nonetheless, we note that an algorithm with optimal asymptotic theoretical guarantees may well be sub-optimal for finite n. Other procedures that are asymptotically optimal may well have much better performance for small to moderate n, so that such analysis, though instructive, does not provide a full picture.

Going back to the functional parameter spaces above, equipped with the L^p norm-distance, $1 \leq p \leq \infty$, if $\mathcal{F}$ is a subset defined by the parameters β and L as in (1.2), it is known [71] that

$$r_n^* = C(\beta, L) \begin{cases} n^{-\frac{\beta}{2\beta+d}}, & p < \infty \\ (n/\log n)^{-\frac{\beta}{2\beta+d}}, & p = \infty \end{cases} \tag{1.4}$$

for $C(\beta, L) > 0$.

1.1.3 Bayesian nonparametrics

Bayesian inference adopts the following point-of-view. Since the parameter is unknown, it is seen as random and one considers the available observations as fixed. A probability distribution Π on the measurable parameter space $(\mathcal{F}, \mathfrak{B})$, called the *prior*, quantifies this uncertainty on the parameter, before data are observed. As above, a model (1.1) gathers the potential distributions of the observations, *given* the value $f \in \mathcal{F}$.

This leads to define a Bayesian model as a probability space

$$(\mathcal{F} \times \mathfrak{X},\ \mathfrak{B} \times \mathfrak{A},\ \Gamma) \tag{1.5}$$

where $\Gamma = \Pi \otimes P_f$ is the joint distribution of $\left(f, \mathbf{X}^{(n)}\right)$. This gives rise to the conditional distribution of $f \mid \mathbf{X}^{(n)}$, called the *posterior distribution* (or simply the *posterior*). In the following, we assume that the model is dominated: there exists a σ-finite measure ν on $\mathfrak{A}$ such that, for any $f \in \mathcal{F}$, $P_f^{(n)}$ is absolutely continuous relatively to ν. This is the case in the density estimation model, with $\nu = \lambda^{\otimes n}$ (as in Sections 2 and 3) and the regression models of Section 5. Under such condition, defining the likelihood of the data $p_f^{(n)}(\mathbf{X}^{(n)}) := dP_f^{(n)}/d\nu(\mathbf{X}^{(n)})$, a version of the posterior distribution is obtained from Bayes' formula, for $X = \mathbf{X}^{(n)}$:

$$\Pi\left[B|X\right] = \frac{\int_B p_f^{(n)}(X) d\Pi(f)}{\int p_f^{(n)}(X) d\Pi(f)}, \qquad B \in \mathfrak{B}. \tag{1.6}$$

An interpretation is that the posterior quantifies the remaining uncertainty on the parameter, once data have been observed. The quantity is always well-defined if the observation $\mathbf{X}^{(n)}$ is sampled from the marginal distribution on $(\mathfrak{X}, \mathfrak{A})$ given by the Bayesian model (1.5). Under a mild condition, namely that the denominator is positive almost surely under $P_{f_0}^{(n)}$, the

posterior distribution $\Pi[\cdot|\, X]$ is also well-defined almost surely under the distribution $P_{f_0}^{(n)}$. This motivates the adoption of a frequentist point-of-view in the use of Bayesian posteriors and their use as tools for inference. After acquisition of the data, the information it contains can hopefully be used to decrease our uncertainty about the parameter f_0, Bayes' rule telling us how to change our 'beliefs' (i.e., the prior) on the parameter.

Going back to the statistical problems mentioned in Section 1.1.1, estimators of f_0 can be deduced. For instance, central aspects such as the posterior mean or the center of balls with large posterior mass (see next section) can be used. It is also conceivable to use samples from the posterior distribution as random estimators. Moreover, these Bayesian methods have an inherent ability to quantify uncertainty (cf. Section 1.3) via their credible sets, i.e. sets with prescribed amount of mass under the posterior. This is a reason for the appeal for Bayesian inference methods: once the posterior distribution has been computed, it contains all the information we need to provide answers to various statistical questions. Bayes methods are then full inferential machines. By contrast, note that typical frequentist methods require to proceed in two steps to perform estimation and uncertainty quantification.

1.1.4 Theoretical analysis of Bayesian posteriors

We recall that one works under the assumptions that X is sampled from $P_{f_0}^{(n)}$, for f_0 fixed, so that the following theory is the one of 'frequentist Bayes'.

Posterior contraction rates

In order to be able to validate Bayesian inferential methods from the frequentist perspective, a natural requirement should be for the posterior to allocate a fair share of its mass to elements of $\mathcal{F}$ that are close to f_0. Equipping the parameter space with a (pseudo-)distance d, an asymptotic analysis of the posterior distribution (see Section 1.1.2) motivates the following definition.

Definition 1 (Posterior contraction rates)**.** *A sequence ε_n is a posterior contraction rate at the parameter f_0 with respect to the metric d if the following convergence holds in $P_{f_0}^{(n)}$-probability*

$$\Pi[f:\ d(f,f_0) \leq M_n\varepsilon_n|\, X] \to 1, \tag{1.7}$$

for every $M_n \to \infty$, as $n \to \infty$.

This essentially means that the posterior concentrates most of its mass on a shrinking ball of radius (almost) ε_n and centered on f_0, and information from the data should eventually overcome the prior in the limit of large n. For the same reason that we seek convergence rates for estimators that hold uniformly on the parameter set, posterior contraction rates should be uniform. However, in the following, we often omit to write it when it is clear that it holds.

Obtaining a posterior contraction rate ε_n at f_0 implies that the center $\tilde{f}_n$ of the smallest ball that contains posterior mass at least $3/4$ satisfies $d\left(\tilde{f}_n, f_0\right) = O_P(\varepsilon_n)$ in $P_{f_0}^{(n)}$-probability, see [8]. Minimax convergence rates (in probability) corresponding to a certain class of smooth functions then act as a lower bound on posterior contraction rates around the true function, assuming it belongs to this given class.

As contraction rates are foremost a feature of the posterior, it seems complicated at first to determine them in situations where the posterior cannot be explicitly computed. However,

since the seminal works of [64, 152], there is by now a general theory for posterior contraction rates in terms of "testing" distances. Those are distances d satisfying the following condition for any $n \geq 1$, $\varepsilon > 0$, $f_0, f_1 \in \mathcal{F}$ such that $d(f_0, f_1) > \varepsilon$, there exists a test $\varphi_n : \mathfrak{X} \to \{0; 1\}$ such that, for some universal constants $\xi, K > 0$,

$$\mathbb{E}_{f_0}^{(n)} \varphi_n \leq e^{-Kn\varepsilon^2}, \qquad \sup_{d(f,f_1)<\xi\varepsilon} \mathbb{E}_f^{(n)} (1 - \varphi_n) \leq e^{-Kn\varepsilon^2}. \tag{1.8}$$

For instance, if we focus on the density estimation model, the Hellinger distance

$$h(f_0, f) = \sqrt{\int \left(f_0^{1/2} - f_0^{1/2}\right)^2 d\lambda}$$

and the L^1-norm distance verify (1.8). Under such condition, the following theorem gives conditions on the prior and the size of the model allowing for the characterisation of posterior contraction rates. The prior mass condition is measured in the Kullback-Leibler divergence $K(f_0; f) = \mathbb{E}_{f_0} \log(f_0/f)$ and the second Kullback-Leibler variation $V(f_0; f) = \mathbb{E}_{f_0} \left(\log(f_0/f) - K(f_0; f)\right)^2$, and the size of the model to the entropy number of a subset, that is the logarithm of the covering number $N(\varepsilon, A, d)$ of a set A which is the minimal number of d-balls of radius ε needed to cover A. Below, we define

$$B_2(f_0, \varepsilon) \coloneqq \left\{f \mid K(f_0, f) \leq \epsilon^2, \ V(f_0, f) \leq \epsilon^2\right\}.$$

Theorem 1 ([64]). *For a distance d on $\mathcal{F}$ satisfying (1.8), if there exists a partition $\mathcal{F} = \mathcal{F}_{n,1} \cup \mathcal{F}_{n,2}$ and $C > 0$ such that*

1. $\Pi(B_2(f_0, \varepsilon_n)) \geq e^{-Cn\varepsilon_n^2}$;
2. $\log N(\xi\varepsilon_n, \ \mathcal{F}_{n,1}, \ d) \leq n\varepsilon_n^2$;
3. $\Pi(\mathcal{F}_{n,2}) \leq e^{-(C+4)n\varepsilon_n^2}$,

then ε_n is a posterior contraction rate for the posterior $\Pi[\cdot \mid X]$.

Informally, this theorem states that, for the posterior to concentrate at a given rate, the prior should not be too wrong in the beginning and put a sufficient amount of mass near the true density f_0. Also, the set $\mathcal{F}_{n,1}$ of limited size should contain the majority of the mass, so that the prior is not too complex. We note that this last point is crucial in the above theorem as it allows for the construction of exponentially consistent test functions [15, 98], an essential element for these arguments. However, while there exist results on lower bounds for posterior contraction rates showing the necessity to have a control on the small ball probability, it is not as clear whether the sieve conditions are necessary (see tempered posteriors below). Intuitively, these conditions ensure that collecting the information from the data is enough to overcome the prior information and obtain a posterior with good concentration properties.

Outside of the canonical metrics satisfying (1.8), it seems much more difficult to obtain generic contraction results in "stronger" distances and these are much sparser in the literature. In particular, when it comes to the supremum norm, we mention the work of [28] that propose method to derive minimax optimal sup-norm rates of convergence, assuming a Hölder regularity of the parameter in density estimation and the Gaussian white noise model for natural families of nonparametric priors and nonconjugate priors, borrowing techniques from semiparametric

statistics. Also, [78] derive adaptive supremum norm rates in the Gaussian white noise model with a sparse prior, the spike and slab prior, as well as rates for more abstract "sieve" priors in various models. In density estimation, supremum norm rates for the classical Pólya tree prior (see Section 2.2.3) were obtained in [29], and adaptive results for a modified prior were proved in [31].

Tempered posteriors

In the above Theorem 1, we may consider Condition 1 as the most important one. We naturally expect that a prior concentrated around the true parameter should give a posterior with good properties. Indeed, with ideas already available in a PAC-Bayesian and Machine Learning context in [35, 179] and then developed in Bayesian nonparametrics in [10, 93, 65], it has been shown that this condition is enough for a modified posterior distribution, the ρ-posterior and some metrics, the Rényi divergences. For $0 < \rho \leq 1$, the *ρ-posterior distribution* is defined similarly to (1.6) but with a fractional likelihood:

$$\Pi_\rho [B|X] = \frac{\int_B p_f^{(n)}(\mathbf{X}^{(n)})^\rho d\Pi(\theta)}{\int p_f^{(n)}(\mathbf{X}^{(n)})^\rho d\Pi(\theta)}, \qquad B \in \mathfrak{B}. \tag{1.9}$$

Chapters 2 and 5 feature results for such (pseudo-)posterior. For $\rho-$posteriors, results are often naturally derived for posterior contraction rates in the Rényi divergence of order β between densities

$$D_\beta(p, q) := \frac{1}{\beta - 1} \log \int p^\beta q^{1-\beta} d\nu.$$

This involves results for the Hellinger distance as well in "properly specified settings" as can be seen from the following inequality

$$D_{1/2}(p, q) \geq 2h^2(p, q).$$

The random probability measures (1.9) are appealing because contraction rates ε_n entirely depend on the prior concentration, namely on the condition that, for some $c > 0$,

$$\Pi \left(f : \ K(f_0; f) < \varepsilon_n^2 \right) \geq e^{-c\varepsilon_n^2}.$$

Compared to Theorem 1, there is no need for the construction of sieves $\mathcal{F}_{n,1}$ and the verification of testing conditions. Therefore, it is possible to conduct a theoretical analysis of a much broader class of priors and models. Notably, condition 3 of Theorem 1 may sometimes impose conditions on the tail of the prior, while the construction of sieves in Condition 2 may be nontrivial to verify as well.

Although the tempering of the posterior typically results in a loss in terms of constants for the convergence rate for fixed ρ, or leads to to credible sets enlarged by roughly $\rho^{-1/2}$ (as one can see e.g. in the simple $\{\mathcal{N}(\theta, 1)^{\otimes n}, \ \theta \in \mathbb{R}\}$ model, where the ρ-posterior for a $\mathcal{N}(0, 1)$ prior on the θ is $\mathcal{N}(\rho n\bar{X}/(1 + \rho n), 1/(1 + \rho n))$, with $\bar{X}$ the empirical mean of the observations), in the nonparametric applications considered in this thesis, their use can be appealing in some cases as this loss is not too detrimental for rates (this loss appears as a constant factor in the rates). However, this may be unsatisfactory for a practical use, with finite amount of available data.

Shape results / Bernstein-von Mises theorems

While the above results are mostly indicative of the speed of convergence of posterior samples to the actual sampling parameter, there is more information than just a rate in the posterior. It is also interesting to investigate the shape of the posterior.

In parametric statistics, in smooth sampling models indexed by a finite-dimensional parameter and if the prior has positive density, the (random) posterior distribution converges in some sense to a normal distribution. When the object of inference is $\theta_0 \in \Theta \subset \mathbb{R}^d$, $d \geq 1$, and $\hat{\theta}_n$ an asymptotically efficient estimator with asymptotic variance proportional to the inverse $I_{\theta_0}^{-1}$ of the Fisher information matrix, the posterior distribution of θ converges as follows

$$\mathcal{L}\left(\sqrt{n}(\theta - \hat{\theta}_n) \,|\, X^{(n)}\right) \rightsquigarrow \mathcal{N}\left(0, I_{\theta_0}^{-1}\right),$$

where convergence is in total variation, in probability under $P_{\theta_0}^{(n)}$. Such result is known as the Bernstein-von Mises (BvM) theorem [98]. From a frequentist point-of-view, this provides a justification of the use of Bayesian procedures as the limit coincides with the efficient frequentist one. Indeed, efficient estimators $\hat{\theta}_n$, such as the maximum likelihood estimator under regularity conditions, satisfy

$$\mathcal{L}\left(\sqrt{n}(\theta_0 - \hat{\theta}_n)\right) \rightsquigarrow \mathcal{N}\left(0, I_{\theta_0}^{-1}\right),$$

when the true parameter is θ_0. Also, since the limiting distribution does not depend on the prior, the accumulating information from the data subsumes the information encoded in the prior.

As for nonparametric models, the obtention of BvM-type theorems is non-straightforward [32, 33]. However, as explained in Section 1.3.1, such results are of particular interest for the construction of confidence sets and the development of nonparametric BvM theorems has a clear appeal for uncertainty quantification purposes. In the Gaussian white noise model [99, 32] Leahu [2011], and in nonparametric regression and density function estimation [33] obtained results in this vein. In these last works, the authors build on a parametric BvM theorem for a finite number of real-valued linear functionals of the infinite-dimensional parameter to obtain a nonparametric BvM theorem. Finally, we point out that these limiting shape results can extend to linear functionals of the parameters as well, e.g., Donsker-type theorem for the cumulative distribution function in the density estimation model (see Theorem 13 of the present document). In Chapter 3, we use this methodology and develop BvM-type results for a particular choice of prior on densities.

1.1.5 Adaptation problems

With the Hölder norm defined as in (1.3), it follows that Hölder balls (1.2) are nested (Lemma 10, [59]),

$$\Sigma(\beta', L) \subset \Sigma(\beta, L), \qquad \beta' \geq \beta, \qquad L \geq 0.$$

Minimax estimation rates over these balls (1.4) show that the estimation problem is easier whenever the target is smoother. Therefore, emphasis should be put on estimators that are able to leverage the higher regularity of the parameter and attain the faster rates of convergence, whenever the true underlying parameter belongs to one of these smaller balls.

As smoother signals renders the inference problems easier, we also expect confidence sets to be more informative under these conditions, that is we expect them to be smaller. In particular,

we would like them to have radii of the same order as the minimax rate of estimation. While adaptive estimators exist for most models, adapting to any regularities, the construction of adaptive confidence sets is a more involved problem and we present some fundamental elements of theory in Section 1.3.2.

Though an overview of methods for adaptation is out of the scope of this introduction, we mention Lepski's method, thresholding estimators as well the natural introduction of layers of hyperpriors to form hierarchical Bayesian priors.

1.2 Some Machine/Statistical Learning methods

In the recent history of learning and inference algorithms, it is often the case that a boost in computing power led to a resurgence in the use of previously impracticable methods.

Notably, with the advent of new computational methods in the recent years, e.g. MCMC, there has been a surge in the use of Bayesian methods. Bayesian counterparts of the original algorithms of Decision Trees and Random Forests (Section 1.2.2), such as BCART and BART, rank amongst the best performers in practice but the theoretical understanding of such achievements is still in its infancy. Another popular class priors on regression surface or probability densities is the Gaussian Processes (GP) priors and their variants (Section 1.2.1). Deep Neural Networks (DNN) is another class that has gained much recent popularity. We briefly draw some links between GPs, forests and DNN in Section 1.2.3.

1.2.1 Gaussian Processes

Gaussian Processes (GP) are stochastic processes that can be viewed as random functions when they are indexed by a set $\mathcal{X}$. They were first used as priors in regression by [88] and [169] and in density estimation by [101] and [102]. Aside from Bayesian nonparametrics, these objects are prominent in Machine Learning as well [138].

Definition 2 (Gaussian process)**.** *A Gaussian process $W = (W_x : x \in \mathcal{X})$ is a stochastic process such that, for any $k \geq 1$ and $x_1, \ldots, x_k \in \mathcal{X}$, the vector $(W_{x_1}, \ldots, W_{x_k})$ is distributed as a Gaussian random vector.*

While this definition takes the viewpoint of GPs as collections of random variables, it is often possible to consider a version of the process with continuous sample paths $x \to W_x$. As a consequence, GPs are often seen as maps $W : \Omega \to \mathbb{B}$, with Ω the underlying probability space and $\mathbb{B}$ a Banach space, e.g., the space of continuous functions equipped with the supremum norm.

These processes are completely characterized by their mean and covariance functions, $\mu : \mathcal{X} \to \mathbb{R}$ and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Most GPs W used as priors in the literature take μ to be identically equal to zero, so that all of the properties of W are determined by K. A popular example of such kernel is $K(s,t) = \min(s,t)$, for $s, t \in [0;1]$, which is associated to the Brownian motion on the unit interval. As a random function, we note that its sample paths are almost surely α-Hölder regular, for $0 \leq \alpha < 1/2$.

A crucial element in the study of these processes as priors on functional parameters is the *reproducing kernel Hilbert space* (RKHS) attached to the Gaussian process (or covariance

kernel). It is defined [164] as the completion $\mathbb{H}$ of the linear space of all functions, $k \geq 1$,

$$x \to \sum_{i=1}^{k} a_i K(s_i, x), \qquad (a_i, s_i) \in \mathbb{R} \times \mathcal{X},\ i = 1, \dots, k. \tag{1.10}$$

in the euclidean norm induced by the inner product

$$\langle \sum_{i=1}^{k} a_i K(s_i, \cdot),\ \sum_{l=1}^{l} b_j K(t_j, \cdot) \rangle_{\mathbb{H}} = \sum_{i=1}^{k} \sum_{j=1}^{l} a_i b_j K\left(s_i, t_j\right).$$

We note that this definition does not depend on the particular representation of elements of the RKHS. The set $\mathbb{H}$ gets its name from the reproducing formula

$$f(x) = \langle f,\ K(x, \cdot) \rangle_{\mathbb{H}}, \qquad x \in \mathcal{X},\ f \in \mathbb{H},$$

which follows for f of the form (1.10) by definition and general $f \in \mathbb{H}$ by completion. This allows us to identify this Hilbert space with a functional space. As an example, the RKHS of the Brownian motion on $[0, 1]$ is the set

$$\left\{ f : [0; 1] \to \mathbb{R},\ f(0) = 0, \int_0^1 f'(t)^2 dt < \infty, f \text{ absolutely continuous} \right\},$$

equipped with the inner product $\langle f,\ g \rangle_{\mathbb{H}} = \int_0^1 f'(t) g'(t) dt$.

A general theory of posterior contraction rates for Gaussian process priors has been obtained in [160, 27, 163]. The RKHS reflects the "geometry" of the Gaussian measure. Building upon the theory on rates presented in Section 1.1.4, these rates ε_n at f_0 in canonical distances are determined as the smallest solution of $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$ where the concentration function at f_0 is

$$\varphi_{f_0}(\varepsilon) = \inf_{h \in \mathbb{H} : \|h - f_0\| \leq \varepsilon} \frac{1}{2} \|h\|_{\mathbb{H}}^2 - \log P\left(\|W\| < \varepsilon\right),$$

and W is a GP with samples in a Banach space equipped with the norm $\|\cdot\|$. The second term, which we write $\varphi_0(\varepsilon)$, is the *small ball probability* and gives a lower bound on contraction rates, independently from f_0. For a Brownian motion viewed as a random element of the space of continuous functions equipped with the supremum norm, $\varphi_0(\varepsilon) \equiv \varepsilon^{-2}$ as $\varepsilon \to 0$. The first term is the *decentering function* and measures the approximation accuracy of the parameter f_0 with elements of the RKHS. A downside of the Brownian motion W for approximation purposes is that it is null at 0. So, we may prefer its released version $Z + W$, $Z \sim \mathcal{N}(0.1)$, to avoid this shortcoming. It has a slightly smaller RKHS, but the small ball probability is of the same order. Then, if $f_0 \in \Sigma(\beta, L)$, $0 < \beta \leq 1$, $L \geq 0$, its decentering function is upper bounded, as $\varepsilon \to 0$, by a multiple of $\varepsilon^{(2-2\beta)/\beta}$. Therefore, a solution of $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$ is $\varepsilon \equiv n^{-(\beta \wedge (1/2))/2}$. This rate is minimax over $\Sigma(\beta, L)$ in Hellinger distance, i.e. proportional to $n^{-\beta/(2\beta+1)}$, for $\beta = 1/2$ only, which matches the regularity of the sample paths. We note that this cannot be improved [27].

Different covariance kernels then lead to different modelling choices and this greatly influences the behaviour of the process as a prior for estimation. Other popular covariance kernels are the square-exponential kernel SqExp, $K(s, t) = e^{-\|s-t\|_2^2}$ with $\|\cdot\|_2$ the euclidean norm and which we study in Chapter 5, and the (α–)Matérn kernel. Both have more regular sample paths than the Brownian motion (if $\alpha > 1/2$ for the Matérn kernel), as the SqExp kernel gives infinitely-differentiable sample paths (and even analytic ones) and the Matérn kernel gives

$\alpha-$regular ones, $\alpha > 0$. We finally note that the RKHS is typically bigger for these smoother GPs, causing the approximation term in the concentration function (which can be understood as a bias) to increase. However, the small ball probability (equivalent to a variance term) decreases. We can then interpret the rate theory of GPs in light of the traditional bias-variance tradeoff from statistical learning.

1.2.2 Tree algorithms and ensembles

Tree-based algorithms find their roots in the "divide-and-conquer" strategy for nonparametric inference. They rely on a recursive partitioning of some relevant space underlying the statistical model, such as the sample space in density estimation or the covariate space in nonparametric regression. They then estimate the parameter locally in each of the obtained subsets, fitting a simple parametric model in those cells, e.g., a constant model. The partitioning pattern can usually be represented in a tree fashion, hence the name of these methods. It is often the case that only binary trees are considered (with binary splits dividing each subset into two smaller ones). Tree ensembles or forests put several trees together.

Tree-based methods and associated ensemble methods rank among the best performers in statistical learning and are, as such, widely used in practice. In nonparametric regression and classification problems, popular versions such as CART (Classification and Regression Trees) [20] and random forests [19] have sparked the interest of numerous researchers. While the former's theoretical behavior in an L^2-loss context is now quite well understood [16, 52, 63] and is still at the heart of recent works [90], the latter have been the object of fewer theoretical results. Indeed, it relies on recursive partitionings of the sample space and piecewise constant predictions, both data-dependent, which renders their analysis quite difficult. To date, research has mainly focused on simplified versions of the original algorithm by Breiman (for which results are scarce [148]) to prove properties such as consistency [3, 11, 48, 82, 89, 118, 150, 168] or asymptotic normality [118, 167]. We refer the interested reader to the review [13] for a more comprehensive account of the topic's literature.

The picture is even more striking with Bayesian versions of these forest estimators, as the theoretical development has just started to emerge. As for the advantages of these Bayesian counterparts, we cite their inherent ability to quantify uncertainty through posterior distributions and their propensity for adaptation. Another noteworthy feature of these methods based on treed partition is that they include the ability to model a nonstationary parameter (where the degree of smoothness is changing over the underlying space) [142]. Popular algorithms are Bayesian CART [40, 49] and BART (Bayesian Additive Regression Trees) [38], the latter being the prototype of Bayesian tree ensemble models. This motivated a myriad of works applying and adapting these methodologies in different domains (e.g., genomic studies [111], credit risk prediction [178], predictions in medical experiments [51], etc.) and to different statistical problems beyond nonparametric regression (e.g., classification [38, 178], variable selection [38, 109], estimation of monotone functions [41], causal inference [75], Poisson processes inference [96], linear varying coefficient models [50], heteroskedasticity [17, 136], etc.). A thorough presentation of the unified framework underlying BART and its extensions appears in [154]. The appeal also finds its roots in such algorithms' competitive practical performance: their great prediction abilities, robustness to ill-specified tuning parameters, and associated efficient posterior computation techniques [72, 73, 134, 135].

At the other end of the spectrum, in spite of their wide application in practice, we are still at the dawn of their theoretical study. We then propose to further the understanding of

such algorithms, in particular via an asymptotic analysis as proposed above in Section 1.1.4. Besides, we should adopt a frequentist point-of-view as presented in 1.1.4 since a reasonable requirement for practical Bayesian procedures is to attain frequentist-optimal posterior rates. It provides a sanity check and is a first step to explain their empirical success, which may be surprising given that these priors are not developed via an expert domain-specific analysis. In very recent years, researchers have come up with results in posterior contraction rate theory for trees and forests in L^2-loss [144, 145], in L^∞-loss [34], and uncertainty quantification [34, 141]. Also, in the context of density estimation, we note that consistency results for BCART-type methods in [113]. For more details, we refer the interested reader to the recent reviews [74, 108].

Finally, we note that [108] provides an informal connection between the BART prior and GPs. According to their heuristic, In the limit of large number of trees in the BART forest, the prior behaves likes a zero-mean GP with explicit covariance kernel. Indeed, tree ensemble methods have a close link with kernel-based methods [149, 3].

Another famous prior in Bayesian nonparametrics that relies on recursive partitioning encoded into a dyadic tree structure is the Pólya tree process. The first related studies date back to the '60s [57, 58, 61, 62] while the name appears for the first time in [97, 116] (it originates in a link with Pólya urns, nicely explained in [65]). As a prior on probability distributions and densities, it has since been used regularly and has recently inspired a new line of research coming up with modified versions [76, 127, 173]. Following this trend, we study priors based on the Pólya Tree construction and their use for inference in Chapters 2 and 3. Lately, posterior contraction rates in L^∞ and Bernstein-von Mises theorems were obtained for Pólya Trees and spike-and-slab Pólya Trees [29, 31].

1.2.3 Neural Networks

Deep learning (DL) is an area of machine learning based on data modelling with complex structures, called Deep Neural Networks (DNN), that compose layers of linear applications and nonlinear transformations based on *activation functions*, with from hundreds to millions of parameters [100, 70]. On each of these layers, numerous of these linear/nonlinear operations (or units) are performed, with the outputs serving as inputs for the subsequent layers. DNNs have proved to be pretty powerful in many problems and enjoyed a surge in computational power to develop at a fast pace in the late 2000s and early 2010s. Their flexibility and capacity to learn feature representations place them among state-of-the-art methods. However, given the complexity of these underlying data structures, for a long time there has been a general belief that no or little theory can be developed for modern deep network architectures. Since tree-based algorithms and Gaussian Processes have a close relationship with DNNs as clarified below, a theoretical analysis of these is valuable to understand why DL works as well.

The compositional structure of DNN motivates analogies with the more recent composition-based Deep Gaussian Processes (DGPs) [44, 56], where multiple GPs are sampled and the output of each one is fed to another layer of GPs. This stochastic process is built from a network of Gaussian processes and takes the form

$$Y_1 \circ \cdots \circ Y_l, \quad l \geq 1,$$

where for $i = 1, \ldots, l$, Y_i is a Gaussian process from $\mathbb{R}^{d_k}$ to $\mathbb{R}^{d_{k+1}}$ ($d_{l+1} = 1$ by convention). This extension of GPs provides a richer class of Bayesian priors, as the sample paths of DGPs can feature more diverse irregularity structures (as in Chapter 5).

In Bayesian neural networks (BNNs), we define a prior distribution on a functional parameter with a neural network whose parameters are drawn randomly, often from a Gaussian or a uniform distribution, and Bayesian inference is done integrating the data likelihood as in (1.6). [123] shows that a neural network with a single hidden layer and i.i.d. parameters converges to a Gaussian process prior, in the limit of a large number of units in the layer, under some conditions ensuring this is well-defined. They also obtain explicit expressions for the limiting covariance kernel. While we may expect a DNN with wide layers, growing altogether, to tend to a DGP, it is actually not true. The limit is a still a GP [115], though with a covariance kernel whose definition is much more involved. In practice, [115] also observed in a simulation study that BNNs and GPs priors with the corresponding choice of the covariance kernel have a similar behaviour. However, [59] remarks that fixing the width of some internal layers in the network composition while the others grow to infinity should lead to a DGP limit. As for traditional DNNs, before the training of the algorithm on the dataset, it has become standard practice to initialize the network parameter randomly, as if it was drawn from a BNN prior. Both BNNs and DGPs offer a probabilistic interpretation and a measure of uncertainty for DL models with their posteriors.

The success of DL algorithms also aroused interest in the potential connections with random forest estimators. First, [151, 172] established that sparse shallow networks can realize these tree-based functions. Then, some recent papers also produced hybrid methods combining upsides of both, e.g., the representation learning properties of DNNs and the computational efficiency of random forests. For instance, we mention *Deep Neural Decision Forests* [91], *conditional networks* [81], *Deep Neural Decision Trees* [175], and *Neural Random Forests* [14], for which consistency in nonparametric regression was proved, starting the theoretical study of these connections.

There is also a surge in the development of the theoretical analysis of DNNs. Nonparametric methods relying on piecewise linear approximations are traditionally suboptimal with signals of smoothness $\alpha > 2$. However DNNs overcome this limitation. Even though sparse DNNs with ReLU activation function results in piecewise linear maps, [147] showed that, in nonparametric regression, near minimax rates for arbitrary smoothness are attained.

1.3 Uncertainty Quantification and confidence sets

In the previous section, we focused on the problem of estimation of a parameter in an infinite-dimensional statistical model. As explained, we should also be interested in the quantification of the uncertainty associated to such estimate. We now describe in a first step the problem of uncertainty quantification (UQ) via confidence sets, before briefly recalling the challenges and limitations in the construction of *adaptive honest confidence sets*, illustrating results in the case where L_p-distances are used to measure the quality of inference.

1.3.1 Construction of confidence sets

In order to do some UQ and account for the inherent uncertainty in our estimation due the randomness of the observations, the typical goal is to produce a sequence of confidence sets $\mathcal{C}_n$, with confidence level $0 < 1 - \gamma \leq 1$, that is

$$\inf_{f \in \mathcal{F}} P_f^{(n)} \left(f \in \mathcal{C}_n \right) \geq 1 - \gamma. \tag{1.11}$$

The asymptotic version has the inequality verified in the limit $n \to \infty$ (possibly a with a limit inferior). These sets allow practitioners to perform inference on the unknown parameter f_0 as their properties are uniform over $\mathcal{F}$.

There are different options to obtain possible values for the parameter and output confidence set. The most basic construction consists in taking the whole set $\mathcal{F}$ as it always contains the true parameter f_0 when the model is well-specified. However, as we explain in the next Section 1.3.2, we should seek the most informative confidence set possible, that is the smallest one. The size of the set quantifies the level of uncertainty of the estimator, and the solution which outputs the whole $\mathcal{F}$ is then not satisfactory. Another possible construction is based on the derivation of the limiting distribution of the estimator under the distribution $P_{f_0}^{(n)}$, when $n \to \infty$. Yet, the problem of deriving this asymptotic distribution is nontrivial in many configurations.

As hinted in previous sections, Bayesian posteriors provide an alternative to these methods. We can indeed define a credible region $\mathcal{C}_n$ with level $1-\gamma$ that verify

$$\Pi[\mathcal{C}_n \,|\, X] \geq 1 - \gamma,$$

or an asymptotic version where the equality is in the limit (possibly the limit inferior) $n \to \infty$.

In the above section, we mentioned analogies between Bayesian algorithms (tree-based and Gaussian processes) with other frequentist methods (kernel-based estimators, Neural Networks, Random Forests). An advantage of the former is that the posterior distribution they provide allows for principled uncertainty quantification procedures. Sections 1.2.1 and 1.2.2 were dedicated to the predictions of the functional parameters that suitable priors could produce once updated. In addition, these distributions also allow the definition of (credible) confidence intervals and give the possibility to generate samples, two features that can prove useful in for UQ purposes.

The question naturally arises whether credible sets can replace confidence sets, i.e., satisfy (1.11) as well, or are misleading in the UQ they provide. We note that the parametric Bernstein-von Mises implies that it is the case in smooth parametric models for quantile credible sets. As for nonparametric inference, in contrast to these, the situation is more delicate and Bayesian credible sets can fail to have correct coverage, i.e., contain the truth with high probability.

Castillo and Nickl [32, 33] obtained nonparametric counterparts of the Bernstein-von Mises theorem in a generic framework, studying as examples the density estimation, nonparametric regression and Gaussian white noise models, which allows them to produce confidence sets from posterior distributions, although for weaker norms than usual L^p $(p \geq 1)$ ones. They also propose a more complex construction to build sets in usual norms by intersecting with sets encoding further qualitative information about the parameter. It allowed them to obtain good frequentist coverage for their sets, coupled with a control of their size in the more usual norms.

We also mention the works of [153, 143] which obtain good coverage properties for credible balls with radius inflated by a blow-up factor, and recover these results adaptively under general qualitative assumptions on the parameter.

1.3.2 Adaptive honest confidence sets

Confidence sets satisfying (1.11) are called *honest* as their coverage is ensured uniformly over the parameter set. Since we should be interested in the sets that are the most informative

possible, a condition on their diameter $|C_n|_d := \sup\{d(f,g) : \ f,g \in C_n\}$ of the form

$$\sup_{f\in\mathcal{F}} P_f\left(|C_n|_d > Lr_n\right) \leq \delta,$$

where r_n is the minimax rate of estimation over $\mathcal{F}$ and $L > 0$, is natural. This condition on the radius is uniform over the parameter set $\mathcal{F}$, but as noted in Section 1.1.5, sets defined via regularity conditions contain parameters with different regularities, e.g. $\mathcal{F} = \Sigma(\beta', L) \cup \Sigma(\beta, L)$, $\beta' > \beta$, in the nonparametric regression model. For simplicity of presentation in this introductory Chapter, we shall consider here only two different regularities, which already illustrates the main ideas. In this case, the minimax rate of estimation $r_n(\beta')$ over $\Sigma(\beta', L)$ is faster than the one, $r_n(\beta)$, on $\Sigma(\beta, L)$. Therefore, we should be interested in adaptive versions of the radius shrinkage condition:

$$\max_{t\in\{\beta',\beta\}} \sup_{f\in\Sigma(t,L)} P_f\left(|C_n|_d > Lr_n(t)\right) \leq \delta.$$

The existence of such sets crucially depends on the geometry induced by the distance d on $\mathcal{F}$. The examples of the L^∞ and the L^2 norm distance illustrate this point [105, 112]: for the former, it is impossible to adapt to any two distinct regularities, while for the latter, adaptation is possible if and only if the two regularities of interest belong to a small interval of the form $[r, 2r]$, $r > 0$.

From an information-theoretical point-of-view, the impossibility to construct adaptive sets when regularities are two different is linked to an eventual discrepancy between the rates of estimation and testing rates in a related testing problem. Therefore, whenever one of these rate is undetermined, it is unclear whether or not one can construct such confidence sets. Indeed, for this reason, there is an array of distances for which the question of UQ in those distances is unanswered.

However, we note that more positive results are available in the literature. These concern weakened programs where the above conditions of honesty and on the radius should be verified on smaller subsets of the parameter sets, defined via additional qualitative assumptions so that the functions regularity can be well estimated, e.g., self-similarity conditions as in Chapter 3.

1.4 Main Questions and outline of the thesis

The present thesis aims at answering some general questions regarding the behaviour of statistical learning and uncertainty quantification procedures.

- Chapter 2: Are there advantages in the use of tree ensembles over a single tree in terms of adaptation to smoothness?
- Chapter 3: Do tree-based methods lead to optimal estimation in terms of the supremum norm? What about uncertainty quantification in terms of confidence bounds?
- Chapter 4: What are the information-theoretic boundaries for uncertainty quantification in terms of Wasserstein losses?
- Chapter 4: Is it always the case that estimators, adapting to any regularities, do not translate straightforwardly into adaptive honest confidence sets?
- Chapter 5: How can Gaussian processes be cast into more general frameworks to tackle more structured parameters?

The following sections provide more insights into these questions.

1.4.1 Bayesian forests

In Chapter 2, we conduct an analysis of some Bayesian forest posteriors, studying their asymptotic properties. Following the discussion on posterior contraction rates in Subsection 1.1.4, priors of choice should be able to attain (or be as close as possible to) Hellinger rates $n^{-\alpha/(2\alpha+1)}$, which are minimax under our assumptions on f_0, at least if $\alpha \leq 1$. Here, we consider the density estimation model from Example 1, with $\mathcal{X} = [0;1)$, and we assume that the sampling density has an arbitrary Hölder regularity $\alpha > 0$.

One limitation of samples from tree posteriors is their lack of smoothness, which diminish their performance as estimators in context where we typically expect the signal to be smooth. Indeed, in nonparametric regression, Bayesian histogram estimators on regular partitions and with independent heights can be shown to be suboptimal for regression functions smoother than Lipschitz-regular, i.e., when $\alpha > 1$ above. Consequently, we may conjecture that the piecewise nature of finite Pólya Tree samples hinders their ability in the estimation of smoother densities as well. However, concerning forest priors, [37] remarks that, in terms of root mean square error, the BART posterior mean (estimated with MCMC samples) fares almost as well as a GP posterior mean on a benchmark simulated regression dataset. This is a seemingly surprising result given that the BART estimates are piecewise constant as well and the true parameter in their study is a smooth function. Recalling the link between BART and GPs mentioned in Section 1.2.2, this suggests that BART samples have a regular underlying structure in spite of their rough piecewise nature.

The 'tree-aggreggation' aspect of BART seems to account for this behaviour. On the other hand, initial work from [3] showed that by aggregating infinitely many randomly shifted tree estimators in a frequentist forest estimator, one can perform smoothing reminiscent of

triangular kernel estimators and adapt to regularities $\alpha \leq 2$ (for regression and for points that are 'far' from the boundary of the covariate space).

In order to shed light on this potential smoothing effect of aggreggation, in Bayesian forests this time, we propose in this chapter a new Discrete Pólya Aggregate (DPA) prior. It is defined as an aggregation of Pólya trees with deterministic shifts and with correlated heights. Informally, the DPA prior on densities is the distribution of

$$f = \frac{1}{q} \sum_{i=0}^{q-1} g_i,$$

where each g_i is distributed as a finite Pólya Tree of depth L, though they are not independent, and with suitable priors on the trees depth L and their number q in the forest.

We demonstrate that such priors, with good hyperparameters, are optimal (up to logarithmic terms) for estimating α-Hölder densities in Hellinger distance for any *arbitrary* $\alpha > 0$. Since the smoothness of the prior is 'quantified' by the depth of the trees, adding a hyperprior on this aspect guarantees adaptivity. The adaptive version of DPA is then optimal for all $\alpha > 0$ simultaneously. The trees in the above aggregation are defined as 'suitable shifts' of the same Pólya Tree sample, so that they are highly 'correlated'.

Firstly, this improves on previous results on tree-based methods [3, 121] as they usually assume that $\alpha \leq 1$, as well as on previous results on random forests models that generally assume $\alpha \leq 2$. Forests were proved to have an advantage over single-tree estimators only on the range $1 < \alpha < 2$. As for Pólya tree-like priors in particular, this brings a partial answer to the question of adaptation and optimality to regularities higher than Lipshitz, a question left open in [29] for the original tree prior. Secondly, another important novelty of these results is that DPA carefully handles the boundary of the unit interval in its definition, while previous results on random forests obtained suboptimal rates at the boundaries when $1 < \alpha \leq 2$. The theory then illustrates further the smoothing benefits of Bayesian forests and their truncations.

We build our results on the general methodology developed around Theorem 1 and we leverage similarities between DPA and spline priors on densities. While [149, 3] linked forests to kernel-based estimators, this link is a further confirmation that ensemble versions of tree-based methods can possess and benefit from hidden regularity structures, in spite of their piecewise-constant nature. Also, although we focus on the density estimation model (a reason being that a goal was to extend results on Pólya Trees in this framework) on the unit interval, the idea is sufficiently general so that it should extend to sample spaces $\mathcal{X}$ in higher dimensions and to other models, such as nonparametric regression (considered in [3]).

1.4.2 Optional Pólya Trees

The results announced in this section were developed in a joint work with Pr. Ismaël Castillo (Sorbonne Université).

Establishing a theory of rates in the supremum norm is all the more appealing as scientists intuitively measure closeness with this distance: when curves are said to be visually close to each other, we interprete this statement with closeness in supremum norm. While the prior proposed in the previous chapter is shown to be adaptive optimal in Hellinger distance, the methodology adopted does not answer the question of supremum norm posterior contraction rates. Theorem 1 does not apply to this stronger distance and this problem is more delicate in general. In addition, as explained in Section 1.1.3, a motivation in the use of Bayesian

approaches is the construction of confidence regions for the parameter. Chapter 3 considers the same density estimation model as Chapter 2 and adopts a Bayesian perspective as well.

[29] obtains results in this direction for the original Pólya Tree construction, showing its nice properties for the estimation of unknown densities. Nonetheless, his construction is non-adaptive as he tuned hyperparameters based on the known regularity $0 < \alpha \leq 1$ of the density, and it remained an open question how the methods could be made adaptive.

Given the link between Pólya trees with dyadic splits in the partition and the Haar wavelet basis, the aforementioned work uses the mutiscale approach introduced in [32, 33] to obtain rates as well as a Bernstein-von Mises theorem. The idea is that a BvM-type theorem can be proved to hold in weaker topologies (from multiscale norms) for some priors, and the resulting credible sets appropriately modified will have asymptotically the correct coverage and optimal size. In a recent work, [31] obtained L^∞-adaptive versions of these theorems for a "spike and slab" version of Pólya Trees. While this relies on an adaptive selection of the wavelet coefficient, this selection is not performed in a tree-fashion and it could be costly in high-dimension.

The original Pólya Tree algorithm involves dyadic complete binary trees, which may be too rigid to expect adaptation. In order to answer the above questions, we consider here a more flexible tree-prior on densities, the Optional Pólya Tree [173], or OPT henceforth: it is a BCART-type hierarchical prior where trees are distributed as a Galton-Watson process (they are no longer complete a.s.) and, given a tree, the random densities are defined in Pólya Tree-like way along the tree. Active coefficients selected in the wavelet basis then depend on the node in the tree.

[34] studied the BCART prior in the Gaussian white noise and nonparametric regression model, in which they obtain Bayesian adaptation for the supremum loss, for estimation and UQ purposes. We remark that the Pólya Tree counterpart of their prior is a specific case of OPT. We pursue their investigation accordingly, in the density estimation model, and adopt the multiscale approach of [32, 33]. As we first prove that the OPT prior is conjugate in this model, we leverage the representation of densities using Haar wavelet expansion to obtain adaptive supremum norm posterior contraction rates, whenever the density has an α-Hölder regularity, $0 < \alpha \leq 1$. We also define a frequentist tree-based estimator from a central aspect of the posterior, the median tree estimator, which is shown to converge at the minimax rate to the true sampling density f_0.

The novelty in this chapter is the validation of tree-based methods in density estimation as a tool for near-optimal uncertainty quantification in terms of the supremum norm. A first answer to questions of estimation and UQ for f_0 is the proposal of an L^∞-ball centered on the median tree estimator which is shown to be, under self-similarity assumptions, an asymptotic credible and confidence set with coverage converging to 1 and an adaptive radius, shrinking at the minimax rate of estimation. A similar construction is proposed for a smooth functional, the cumulative distribution function associated to f_0, this time with exact (credible and confidence) coverage $1 - \gamma$ and without self-similarity assumptions. To obtain exact asymptotic coverage for f_0, we derive an adaptive BvM theorem in a weak multiscale distance, along the lines of [139], which is a key element towards the obtention of confidence bands. The intersection of the supremum norm ball centred on the median tree with a multiscale credible band of level $1 - \gamma$ is shown to have exact asymptotic coverage as well as adaptive supnorm radius.

The conjugacy property of the OPT prior in the density estimation model and its sequential definition is also appealing from a practical point of view. Exact sampling from the posterior

turns out to be possible using the recursive nature of the tree partitioning, and we conduct a numerical study on synthetic dataset to illustrate the proposed methodologies for estimation and UQ and their features, including regularity adaptation, spatial adaptation [142] and the limitation to Lipschitz-regularity for methods based on single trees (as underlined in Chapter 2).

1.4.3 Uncertainty Quantification with Wasserstein distances

The results announced in this section were developed in a joint work with Neil Deo (Cambridge University).

In Chapter 4, we focus on the question of the construction of adaptive honest confidence sets for f_0 as presented in Section 1.3. As explained, the theory underlying the possibility for such sets to adapt to different regularities depends on the geometry induced by the distance used for inference. In particular, the (non-)existence of adaptive sets is partly determined by the minimax estimation rates over the smoothness classes of interest and rates in some related testing problems. Whenever one of these rates is unknown, the question of existence remains unanswered. We consider here the density estimation model with the $d-$dimensional torus as sample space $\mathcal{X} = \mathbb{T}^d$ and assume that the sampling density f_0 has a regularity $s > 0$ (measured on the Besov scale of smoothness, see Section 4.5).

This work focuses on the case of the Wasserstein distance, W_p, $1 \leq p \leq 2$, between densities defined as

$$W_p(f,g) := \inf_{\pi \in \Pi(P_f, P_g)} \left(\int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi(x,y) \right)^{1/p},$$

with the infimum ranging over the set $\Pi(P_f, P_g)$ of measures on $\mathcal{X} \times \mathcal{X}$ with marginals P_f and P_g, that are the distributions on $\mathcal{X}$ with Lebesgue densities f and g respectively. While this distance finds its roots in Optimal Transport problems [120, 87], it recently became quite popular in several fields such as optimization [131], computer vision [124] or Machine Learning with the development of Wasserstein GANs [2, 12].

It also has appears in Statistics in several ways, as a tool for inference. In particular, [171] determined the minimax rates of estimation for Besov-regular densities that are lower and upper bounded by positive constants. They showed that, under these conditions, the Wasserstein distance compares with weak Besov norm-distances of negative smoothness -1, extending the classical Kantorovich-Rubinstein duality formula applied to W_1. As a consequence, the obtained convergence rates $n^{-\frac{s+1}{2s+d}}$ are faster than the usual L_p–rates. However, neither the question of testing rates in Wasserstein distance nor the construction of honest adaptive confidence sets were tackled.

Leveraging this comparison with a negative Besov norm, we obtain lower bounds on the testing rates in arbitrary dimension d, once again much faster than the corresponding L_p ones, which allows us to discard regularity values for which adaptation is not feasible. When it comes to the positive results, in all the situations that have not been ruled out by the preceding argument, we provide a construction of confidence sets. We define these as balls in a sample-splitting scheme, first constructing an adaptive minimax estimators as a center and using the method of risk-estimation via U-statistics to determine the radius. For this last point, we indeed use continuous embeddings of negative Besov spaces in some Hilbert space so that it is possible to estimate the risk at a very fast rate, allowing for adaptation.

Our results features a surprising new phenomenon in view of the classical theory for L^p distances: in small dimension, it is possible to adapt to any regularities. In higher dimensions, adaptation is possible if and only if regularities of interest are close enough, lying in window that is still (significantly) larger than the one prescribed for L^p adaptation. This window takes the form $\left[r, \frac{2d-4}{d-4}r + \frac{d}{d-4}\right]$, as opposed to $[r, rp/(p-1)]$ (L^p distance, $p \geq 1$). We underline that our findings are not limited to Wasserstein distances: similar results can be proved for any Besov norm-distances of negative smoothness, granted that Hilbert space embeddings are possible for the positive results. For the negative results, it transpires that fewer configurations are ruled out by the theory because, while estimation and testing rates both accelerates, the later becomes much faster. Regularities $r < s$ to which adaptation is possible simultaneously are the ones such that $n^{-\frac{r+1}{2r+d/2}} \geq n^{-\frac{s+1}{2s+d}}$. Since this effect should also happens for other weak negative distances, an analogous phenomena should take place.

1.4.4 Deep Horseshoe Gaussian Processes

The following elements are the focus of an ongoing project in collaboration with Pr. Ismaël Castillo (Sorbonne Université).

In the last chapter, we consider the nonparametric regression model from Example 2, where we observe n i.i.d, variables $(X_i, Y_i) \in [-1,1]^d \times \mathbb{R}$ and

$$Y_i = f_0(X_i) + \varepsilon_i, \ i = 1, \dots, n,$$

with $i.i.d.$ noise variables $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ independent from the design X_i. The parameter to be recovered there is the regression function f_0. We consider the case where f_0 veryfies an additional compositional structural assumption. We place ourself under the framework of [147, 59] which assume that $f_0 = h_q \circ \cdot \circ h_0$ is the composition of $q+1$ applications. Each of the component $h_i : [-1,1]^{d_i} \to [-1,1]^{d_{i+1}}$ is assumed to be β_i-Hölder and to depend on a small set of variables $t_i \leq d_i$ only.

On the one hand, while Gaussian processes proved to be competitive in practice for a wide range of tasks, theoretical guarantees are nonetheless scarce when it comes to the more current contemporary applications such as high-dimensional inference, where the signal has a low-dimensional structure. Hierarchical extensions of GPs are usually adopted to tackle these situations, performing a rescaling and a stochastic variable selection that allows for an adaptation to the smoothness and the sparsity of the regression function. [176] propose for instance an add-GP prior with a Bernoulli selection of active variables and a conditional GP prior on the selected subset. On the other hand, the presence of a compositional structure in the regression function motivates the study of DGP priors. [59] proposes to place a prior on this latent structure, represented as a network graph with edges representing the selected active variables, and then, conditionally on this graph, they draw the different GPs in the composition. They also condition individual GPs to be 'smooth' enough in order to regularize the composition and avoid paths with wild behaviour detrimental to the estimation goal.

In both situations, we remark that some sort of "hard" selection of variables is performed through a hyperprior on the discrete subset of variables, in a hierarchical prior, ahead of the draw of the GPs. We propose to replace this with a "soft" version of this selection, using the ARD (Automatic Relevance Determination) kernel, akin to the SqExp kernel,

$$K(s,t) = e^{-\sum_{i=1}^{d} w_i (s_i - t_i)^2}.$$

Already used in the original paper on DGP [44], it puts different weight w_i for each latent dimension with the aim to "switch off" irrelevant dimensions with small weights. In this paper, the weights are determined via a variational procedure. We adopt here a fully Bayesian method and put a hyperprior on these variables. Inspired by the Horseshoe prior [26, 159], the weights then have a continuous distribution, with a pole at 0 and sufficiently large tails to simultaneously perform the scaling and the stochastic variable selection. The beneficial effect of the Horseshoe prior in this context is that it allows for the simultaneous adaptation to the regularity and the low-dimensional structure of the parameter.

We demonstrate that the ρ-posterior, for this Horseshoe GP prior and the Deep Horsehoe GP under structural assumptions, achieves minimax contraction rates. For the deep prior, we note that it replaces the sparse structural graph with a full graph with weighted edges, which simplifies the definition of the prior as opposed to [59] and should be more convenient for simulation purposes.

CHAPTER 2

Smoothing from Bayesian forests

Recently, S. Arlot and R. Genuer have shown that a random forest model outperforms its single-tree counterpart in estimating $\alpha-$Hölder functions, $1 \leq \alpha \leq 2$. This backs up the idea that ensembles of tree estimators are smoother estimators than single trees. On the other hand, most positive optimality results on Bayesian tree-based methods assume that $\alpha \leq 1$. Naturally, one wonders whether Bayesian counterparts of forest estimators are optimal on smoother classes, just as observed with frequentist estimators for $\alpha \leq 2$. We focus on density estimation and introduce an ensemble estimator from the classical (truncated) Pólya tree construction in Bayesian nonparametrics. Inspired by the work mentioned above, the resulting Bayesian forest estimator is shown to lead to optimal posterior contraction rates, up to logarithmic terms, for the Hellinger and L^1 distances on probability density functions on $[0;1)$ for arbitrary Hölder regularity $\alpha > 0$. This improves upon previous results for constructions related to the Pólya tree prior, whose optimality was only proven when $\alpha \leq 1$. Also, by adding a hyperprior on the trees' depth, we obtain an adaptive version of the prior that does not require α to be specified to attain optimality.

Table of Contents

2.1 Introduction.

A significant shortcoming of the positive results of (almost-)optimal performances for tree-based methods (ensemble versions or not) is that they typically assume that the signal has limited regularity. Indeed, they study functional parameters that either lie in between step-functions [158] and Lipschitz applications [29, 31] or belong to an additive model with Lipschitz components [145]. This comes from the fact that tree-based partitioning induces piecewise-constant estimators, which are usually too rough for efficient inference of smooth applications. Nevertheless, it has been noted that the aggregation of individual estimates may have a "smoothing" effect. This idea is already present in Breiman's bagging [21]. It thus seems conceivable that a forest, i.e., a tree ensemble, may be more regular and enjoy optimal rates with even more restrictive smoothness assumptions. A few years ago, Arlot & Genuer [3] indeed showed that this could be the case in a regression setting. Their method is described in more detail in Section 2.2. In addition, some works made links between random forests and kernel estimators [149, 3]. This similarity was also recently established by [128] in the context of another tree ensemble model based on the Mondrian process [146], the Mondrian Forest [95]. The forest models developed there are such that the random construction of single trees is independent of the observed data. As mentioned above, such simplification has proved fruitful to come up with theoretical results. These are commonly known as *Purely random forests*. They are also interesting for the study of Bayesian tree ensemble methods in that they too rely on the specification of a probability distribution on sample space recursive partitionings.

The present work builds upon ideas from [3]. The authors developed a *Purely random forest* model that attains minimax convergence rates on the class of twice differentiable functions on $[0; 1]$, for a modified L^2-loss (excluding points near the interval frontier). However, the related single-tree estimators are shown to be optimal only up to once-differentiable regularity. In the following, we will see how it is possible to adapt their aggregation of trees to forests to Pólya trees. We introduce a new prior on probability density functions, the Discrete Pólya Aggregate (DPA) prior, which is a toy random forest incorporating an aggregation step in its definition. Our contribution is threefold. First, we prove that the aggregation in DPA induces smoothing: the corresponding posterior distribution attains the optimal (up to log terms) Hellinger rates on classes of densities of arbitrary Hölder regularity α (no upper bound). Note that [3] and [121] achieve this only up to $\alpha = 2$. We demonstrate this smoothing through a link between DPA and spline densities. Second, our construction is adaptive: our results hold without knowledge of the regularity parameter through a hyperprior on the tree depth. Furthermore, we show how to handle smoothing at the domain's frontier by slightly modifying the prior close to the edges. Third, the DPA prior can be seen as a possible 'way' to smooth PT priors, a question left open in [29, 31], at least here in terms of Hellinger rates. These

results highlight the benefits that ensemble methods can have in Bayesian nonparametrics, smoothing the estimator to attain optimality on a broader class of problems and allowing adaptation. It is worth mentioning that [107] showed that a Bayesian forest made of 'smooth' decision trees adapts to high regularities in a regression setting. However, their individual 'tree predictors' are already smooth, whereas we work with 'hard' histogram-tree predictors, as in original random forests.

This chapter outline is as follows: Section 2.2 introduces our study framework and the aggregation ideas from [3] before presenting the DPA prior. Then, in Section 2.3, we expound on our results on the DPA posterior and other ones on priors inspired by the study of DPA. After a short discussion, the article body ends with the proofs of these results in Section 2.5. Selected proofs, helpful lemmas and elements used to derive our main results and a short numerical study illustrating our theoretical analysis are deferred to Section 2.6.

2.2 Aggregation of a Pólya Tree.

2.2.1 Framework.

Let's elaborate on the problem at hand, which is the one of density estimation. Below, P_f is the probability distribution on $\Omega := [0;1)$ with density f w.r.t. Lebesgue measure λ. In a full Bayesian framework, one specifies a distribution on a pair $(X^{(n)}, f)$, determined by a prior on probability densities f, denoted Π, and the conditional distribution $X^{(n)} \,|\, f \sim P_f^{\otimes n}$. From this, one obtains the posterior distribution, denoted $\Pi[\cdot \,|\, X]$, omitting the superscript for conciseness. We adopt a frequentist point of view in the analysis of the Bayesian procedure. Indeed, we assume that X follows the distribution $P_{f_0}^{\otimes n}$ for a given f_0 instead of the marginal distribution of the pair (X, f). Accordingly, we are interested in the asymptotic behaviour of $\Pi[\cdot \,|\, X]$ under such conditions. In this chapter, we introduce priors on probability densities such that the associated posteriors concentrate their masses on balls with center f_0 at optimal rates (up to logarithmic factors).

A central assumption for our subsequent analysis is that f_0 is α-Hölderian ($\alpha > 0$), i.e., it belongs to the Hölder class, with $\lfloor \alpha \rfloor$ the biggest integer strictly smaller than α,

$$\Sigma(\alpha, [0,1)) := \left\{ f : [0,1) \mapsto \mathbb{R} \;\middle|\; ||f||_{\Sigma(\alpha)} := \sup_{x \neq y} \frac{|f^{(\lfloor \alpha \rfloor)}(x) - f^{(\lfloor \alpha \rfloor)}(y)|}{|x-y|^{\alpha - \lfloor \alpha \rfloor}} < +\infty \right\}.$$

In the following, we write, for real positive sequences u_n, v_n, $u_n \lesssim v_n$ whenever there exists a constant $C > 0$ independent of n such that for any n large enough, $u_n \leq C v_n$ ($\gtrsim$ is defined likewise). Also, if $u_n \lesssim v_n$ and $u_n \gtrsim v_n$, we write $u_n \asymp v_n$, while $u_n \propto v_n$ means that there exists a constant C such that $u_n = C v_n$. When comparing two quantities $a, b \in \mathbb{R}$, we write $a \vee b := \max(a,b)$ and $a \wedge b := \min(a,b)$. For random variables X and Y, $X \perp\!\!\!\perp Y$ means independence. $\mathbb{E}_f$ denotes the expectation under $P_f^{\otimes n}$, as there will no ambiguity on n in the following. Also, as pointed out, we denote $\lfloor a \rfloor$ the greater integer strictly smaller than a. As for the usual floor operator, it is written $\lfloor \cdot \rfloor_{\mathsf{f}}$. For real univariate maps f, g, $f * g$ is their standard convolution. The n-dimensional unit simplex is

$$S^n := \left\{ (x_1, \cdots, x_n) \in \mathbb{R}^n \,|\, x_i \geq 0, \sum_{i=1}^n x_i = 1 \right\}$$

and the ϵ-covering number $N(\epsilon, A, d)$, for some $\epsilon > 0$ and A a subset of metric space (V, d), is the minimum number of ϵ-balls with centers in V needed to cover A. $\|\cdot\|_1$ is the $L^1(\Omega)$ norm and h is the Hellinger distance.

2.2.2 Smoothing of frequentist forest estimators.

Since they entail piecewise-constant estimators with independent heights, inference methods that rely on single-tree constructions are usually limited in their performance. They are generally suboptimal on balls of Hölder classes with regularity $\alpha > 1$. Nonetheless, there is hope that their ensemble methods are less prone to such problems and better suited for the inference of smoother parameters. In this section, we discuss a toy model from [3] which confirms this intuition.

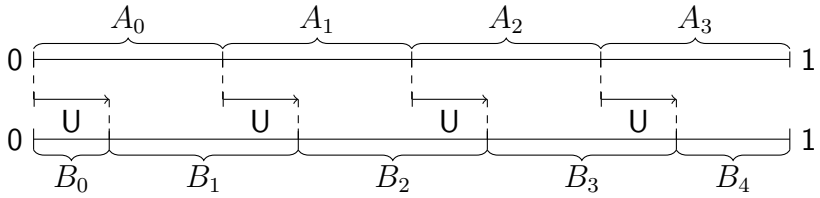


Figure 2.1: Random shift of a regular partition.

Let's assume that we are faced with estimating some map $f_0 \colon [0;1) \to \mathbb{R}$ (let it be in density estimation, regression, Gaussian white noise problem, etc.). Tree-based methods build upon a recursive partitioning of the interval $[0;1)$ so that the estimator is piecewise constant on this given partition. Some data average typically defines the values taken on each cell. Conversely, any partition in subintervals can be represented by a binary tree. Purely random forests then rely on the aggregation of random piecewise constant estimators, the cells of the partitions being random and independent of the observed data.

A particular toy distribution $\mathcal{P}_{\text{toy}}$ on partitions presented in [3] is sketched in Figure 2.1. Given $k \in \mathbb{N}^*$, the partition $\mathbb{U}$ is defined starting with a regular partition of step k^{-1} whose breakpoints are shifted to the right by U/k, with $U \sim \mathcal{U}[0;1]$. A tree estimator averages the data on the partition $P \sim \mathcal{P}_{\text{toy}}$ while a q-forest, $q \geq 1$, is itself the average of q tree estimators corresponding to the i.i.d. partitions $P_i \sim \mathcal{P}_{\text{toy}}$, $i = 1, \dots, q$. In the context of nonparametric regression with a modified L^2-loss that only takes into account points far enough from the frontier of $[0;1]$, [3] shows that forests with a sufficient amount of trees and a well-chosen k attain optimal convergence rates in the estimation of twice differentiable functions. An intuition for this result is that, in the limit $q \to \infty$, the forests actually mimic a triangular kernel estimator (see Proposition 4 from [3]). On the other hand, even with an optimal k, single-tree estimators cannot do better than usual histogram estimators and are optimal only when the function f_0 has at best Lipschitz regularity. While the approximation error of a tree is controlled by the L^2-projection of f_0 on a linear space of piecewise-constant maps, forests' one is controlled by the average of such projections on different spaces (the cells varying between spaces). It then appears that aggregation allows forests to borrow strength from a smoother object that enjoys nice approximation properties, bringing about the estimator's smoothing.

However, aggregating once still only allows the obtention of minimax convergence rates corresponding to twice differentiable regularity at most. It is unclear how the idea could be pushed further to adapt to arbitrary regularities in this context. Below, we see that it is possible to do so with Bayesian estimators.

2.2.3 The DPA prior.

Since we are focusing on density estimation, it is sensible to delve into the Pólya Tree prior, and more particularly into its finite version where the tree is truncated at a given depth. We talk about the Truncated Pólya Tree (TPT) prior to refer to the distribution on probability density functions defined in the following paragraph.

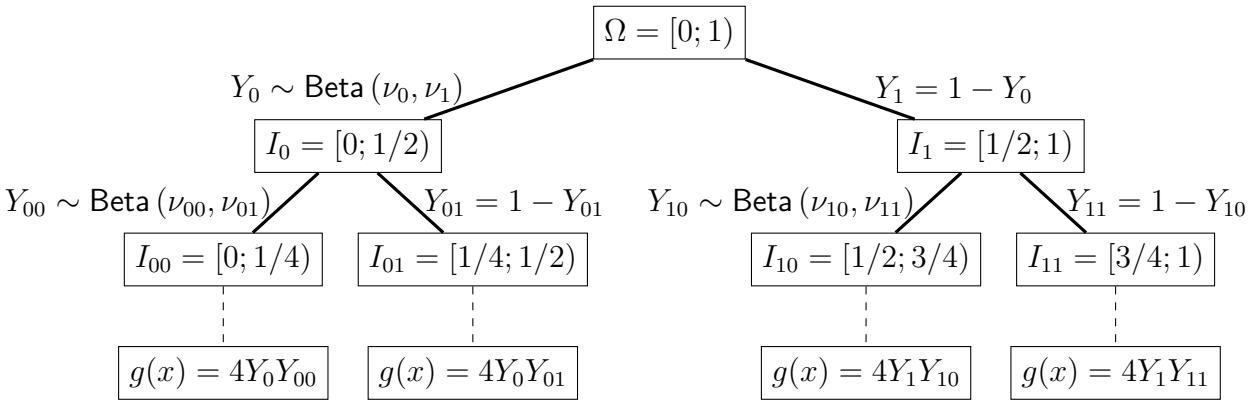


Figure 2.2: Truncated Pólya Tree at depth $L = 2$.

We introduce some notation and illustrate the construction of the TPT prior in Figure 2.2 (see also Chapters 3.5-3.7 of [65]). For any $l > 0$ and $0 \leq k < 2^l$, the dyadic number $r = k2^{-l}$ has a binary expansion $0.\kappa_1 \dots \kappa_l$ (possibly padded with 0's on the right) with $\kappa_i \in \{0, 1\}$ for $i = 1, \dots, l$, i.e., $r = \sum_{j=1}^{l} \kappa_j 2^{-j}$. Reciprocally, any $\kappa \in \{0, 1\}^l$ with $l > 0$ corresponds to the binary expansion of a dyadic number of the form $k2^{-l}$ for some $0 \leq k < 2^l$. Consequently, we write $\kappa(l, k) \in \{0, 1\}^l$ the sequence of length $|\kappa(l, k)| = l$ corresponding to the dyadic $k2^{-l}$ (by convention $\kappa(0, 0) = \varnothing$). Then, for $0 < i \leq l$ and $\kappa = \kappa_1 \dots \kappa_l \in \{0; 1\}^l$, $\kappa^{[i]} := \kappa_1 \dots \kappa_i$. For any pair $(l, k) \in \mathbb{N} \times \mathbb{Z}$, we introduce $I_{l,k} := \left[\frac{k}{2^l}, \frac{k+1}{2^l}\right)$ and, whenever $l \geq 0, 0 \leq k < 2^l$, $I_{\kappa(l,k)} := I_{l,k}$. One sees that $I_\kappa = I_{\kappa 0} \cup I_{\kappa 1}$ and $\Omega = \cup_{\kappa:\ |\kappa|=l} I_\kappa$ so that the union of the sets of the form I_κ defines a recursive partitioning of the unit interval. In Figure 2.2, one sees that this partitioning consists of splitting each interval in its midpoint from one level to another. We refer to $I_{\kappa 0}$ (resp. $I_{\kappa 1}$) as the left (resp. right) child of I_κ. Therefore, κ encodes the sequence of partitioned sets from Ω to I_κ according to this relationship.

Then, for $L \in \mathbb{N}^*$ and a set of positive real parameters $\mathcal{A} = \left\{\upsilon_\kappa, \kappa \in \cup_{l=1}^{L} \{0, 1\}^l\right\}$, we say that the Lebesgue probability density g follows a Truncated Pólya Tree distribution $\text{TPT}_L(\mathcal{A})$ for the above recursive partitioning scheme if there exist random variables $0 \leq Y_\kappa \leq 1$ for any $0 < |\kappa| \leq L$ such that

- the variables $Y_{\kappa 0}$ with $0 \leq |\kappa| \leq L - 1$ are mutually independent and $Y_{\kappa 0} \sim \text{Beta}(\upsilon_{\kappa 0}, \upsilon_{\kappa 1})$,
- $Y_{\kappa 1} = 1 - Y_{\kappa 0}$ for $0 \leq |\kappa| \leq L - 1$,
- $\forall \kappa$ s.t. $|\kappa| = L,\ \forall x \in I_\kappa,\ g(x) = 2^L \prod_{l=1}^{L} Y_{\kappa^{[l]}}$.

The link between this construction and dyadic trees is illustrated in Figure 2.2 where we see that L defines the depth of a tree that encodes a partition of Ω. One sees that for any

sequence κ as above, $Y_{\kappa 0} = P_g(I_{\kappa 0}) / P_g(I_\kappa) = P_g(I_{\kappa 0} \,|I_\kappa)$. Subsequently, we always assume for simplicity that the parameters v_κ in the set $\mathcal{A}$ verify $v_\kappa = a_{|\kappa|}$, so that we rather write $\mathcal{A} = (a_l)_{0<l\leq L}$.

Setting $H_{L,i} \coloneqq 2^L \mathbb{1}_{I_{L,i}}$, a probability density g defined as above can be written

$$g = \sum_{i=0}^{2^L-1} \Theta_i H_{L,i}, \qquad \Theta_i = \prod_{l=1}^{L} Y_{\kappa\left(l,\lfloor i 2^{l-L} \rfloor_{\mathfrak{f}}\right)}, \tag{2.1}$$

so that it belongs almost surely to

$$C_L \coloneqq \left\{ h : [0,1) \mapsto \mathbb{R}^+ \;\middle|\; \int h = 1,\ h \text{ is constant on } I_\kappa, |\kappa| = L \right\}.$$

As pointed out before, the elements of C_L are too rough to approximate efficiently smooth applications, so we would like to leverage ideas developed in the last section to obtain "smoother" prior samples.

The TPT prior defined above has samples that are piecewise constant on some dyadic partition of Ω. However, it would be possible to develop a similar prior so that the samples are piecewise constant on a different partition. Figure 2.3 illustrates such construction in the case of a dyadic partition shifted by some quantity S, just like in Section 2.2.2. We point out a slight difference from what we described in Section 2.2.2: to be encoded in a complete binary tree of depth $l \geq 0$, the recursive partitions need to have 2^l elements at level l, while a shifted dyadic partition has $2^l + 1$ elements (see Figure 2.1). The way to go here is to merge the external cells, corresponding to sets B_0 et B_4 in Figure 2.1 or $[0; S)$ and $[3/4 + S; 1)$ in Figure 2.3. Following the ideas from Section 2.2.2, for some $L > 0$, we could then define a new prior whose samples are the averages of $q \geq 0$ independent maps. Each of these maps would follow an independent TPT prior with depth L, with their underlying partitions that are dyadic partitions shifted by q independent uniform random variables.

As we will see in the following sections, the priors must allocate some of their mass to subsets of limited complexity to obtain posterior contraction rates. It is not the case of the construction we just proposed, as the samples belong to a functional class that is too rich, and the prior mass is overly spread out. Consequently, we need to impose some correlation between the q tree maps. We propose a modified prior in which the aggregated trees are not independent and the shifts of their dyadic partitions are deterministic. Indeed, as we do not want the prior to have its samples that are much more complex than the ones of the TPT distribution (for a given $L > 0$), we would like it to involve just as many Y's Beta random variables.

First, for $f : \mathbb{R} \to \mathbb{R}$, $q \in \mathbb{N}^*$ and $s > 0$, we define a finite aggregation step as the effect of the map $f \to f^1_{q,s}$ defined by

$$\begin{aligned} f^1_{q,s} &: \mathbb{R} \to \mathbb{R} \\ & x \mapsto \frac{1}{q} \sum_{i=0}^{q-1} f\left(x - \frac{is}{q}\right) \end{aligned} \tag{2.2}$$

for any application $f : \mathbb{R} \to \mathbb{R}$. If f is the 1-periodic extension of a TPT_L sample and $s = 2^{-L}$, $L > 0$, the restriction to $[0; 1)$ of $f^1_{q,s}$ is the aggregation of $q \geq 0$ piecewise constant maps as described above. More precisely, it is the aggregation on $[0; 1)$ of q maps constructed in the same way as in Figure 2.3, and that share the same values (the Y's variable are identical) on

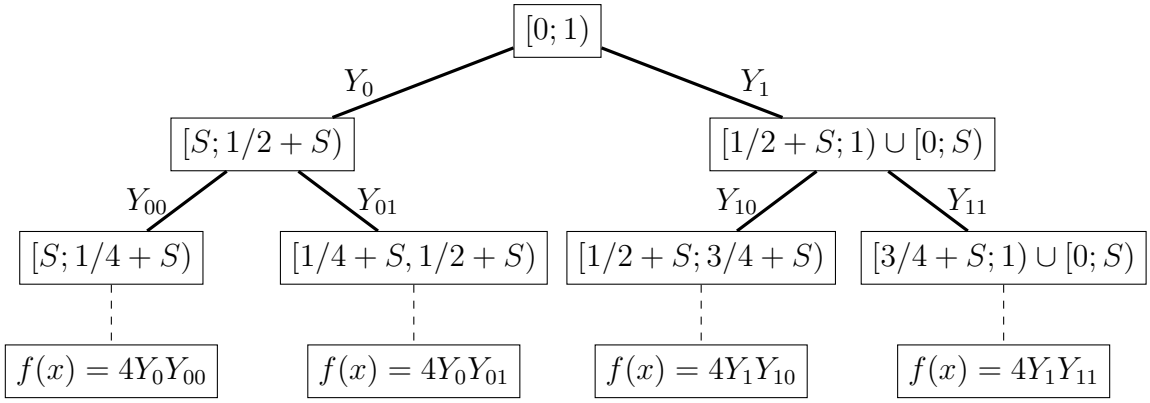


Figure 2.3: Shifted Truncated Pólya Tree at depth $L=2$, with shift S.

their respective underlying partitions, which are dyadic partitions shifted by $S_i = iq^{-1}2^{-L}$ for $i = 0, \dots, q-1$. The resulting map could then be viewed as the sample of a simplified prior on forests, based on TPTs, which is the pushforward measure of the TPT_L measure by the 'aggregating' map (2.2). However, this restriction on $[0;1)$ of a 1-periodic map is 'cyclical' over the frontier of the interval, as most of the trees have the same value near 0 and 1 (cf. Figure 2.3). We discuss this after the complete definition of our new priors.

Before going further, let's discuss some aspects of our aggregation scheme and the definition of the 'Pólya forest' above. It has the benefit that it can be generalized to higher aggregation orders via the following recurrence relationship, which involves some weights depending on the degree of aggregation $k \in \mathbb{N}^*$:

$$f_{q,s}^{k+1}(\cdot) := \left(f_{q,s}^{k}\right)_{q,s}^{1}(\cdot) = \frac{1}{(q-1)(k+1)+1} \sum_{i=0}^{(q-1)(k+1)} \sum_{\substack{(j_1,\dots,j_{k+1})\in[\![0;q-1]\!]^{k+1},\\ j_1+\cdots+j_{k+1}=i}} f\left(\cdot - \frac{is}{q}\right).$$

Aggregating k times, we obtain "forests of forest", which are more general forests of weighted trees, with non-uniform weights. To see the actual effect of this operation, it is useful to have a look at what is happening in the limit $q \to \infty$ (i.e. for an "infinite forest"). If f is Riemann integrable on any interval of length s, then letting $q \to \infty$ results in $f_{q,s}^1$ converging pointwise to

$$\begin{aligned} f_{\infty,s}^1 \colon \mathbb{R} &\to \mathbb{R} \\ x &\mapsto s^{-1} \int_{x-s}^{x} f(t)dt = s^{-1}(\chi_s * f)(x) \end{aligned}$$

where $\chi_s(t) := \mathbb{1}_{[0;1]}(t/s)$. For simplicity, we write $\chi := \chi_1$. This defines a continuous aggregation that can also be iterated, for $l \in \mathbb{N}^*$,

$$f_{\infty,s}^{l+1} := \left(f_{\infty,s}^{l}\right)_{\infty,s}^{1} = s^{-(l+1)}\chi_s^{*(l+1)} * f \tag{2.3}$$

with $\chi_s^{*(l+1)}$ the $(l+1)$-th iterated convolution of χ_s with itself. Such an "infinite forest" (if f is a tree function) is the continuous aggregation of a continuum of 'tree' maps, possibly with non-uniform weights for higher degrees of aggregation. Besides, we take the convention $f_{q,s}^0 := f_{\infty,s}^0 := f$. Thanks to Lemma 3, for $m \geq 1$, we also have the more explicit formula for the continuous aggregation

$$f^m_{\infty,s}(\cdot) = s^{-1}\ \chi^{*m}(\cdot/s) * f. \tag{2.4}$$

As we will see later, the weighted aggregation of trees allows obtaining even smoother forests. Finally, setting, for $0 < i \leq m+1$,

$$\omega_{m,i} := \int_0^i \chi^{*(m+1)}(t)\, dt, \tag{2.5}$$

we define our new prior, the Discrete Pólya Aggregation (DPA) distribution. Fixing $m \in \mathbb{N}$, $L \in \mathbb{N}^*$ such that $2^{L-1} > m$, $q \in \mathbb{N}^*$, $U > 0$ and a set of hyperparameters $\mathcal{A}$, the samples of DPA$(m, L, q, \mathcal{A}, U)$ are generated sequentially as follows:

1. Trees definition and handling of the frontier

   a) Draw g such that $g \sim TPT_L(\mathcal{A})$. One writes

   $$g(\cdot) = \sum_{i=0}^{2^L-1} \Theta_i H_{Li}$$

   for some sequence $(\Theta_i)_{i=0}^{2^L-1}$ whose elements are positive and sum up to 1.

   b) Given Θ_i, $0 \leq i \leq 2^L - 1$,

   $$\theta_i = \begin{cases} \Theta_i & \text{if } 0 \leq i \leq 2^L - m - 1, \\ v_i \sim \mathcal{U}\left[0 \vee \frac{\Theta_i - (1-\omega_{m,2^L-i})U}{\omega_{m,2^L-i}}; U \wedge \frac{\Theta_i}{\omega_{m,2^L-i}}\right] & \text{if } 2^L - m \leq i \leq 2^L - 1, \\ \theta_i = \frac{\Theta_{i-m} - \omega_{m,2^L+m-i} v_{i-m}}{1-\omega_{m,2^L+m-i}} & \text{if } 2^L \leq i \leq 2^L + m - 1 \end{cases}$$

   where the uniform variables above are mutually independent.

   c) Define the $(2^L + m)$-periodic sequence $(u_i)_{i\in\mathbb{Z}}$ such that $u_i = \theta_j$, $j \equiv i \mod 2^L + m$.

2. Aggregation

   Set

   $$f = \sum_{i\in\mathbb{Z}} u_i H_{L,i}$$

   as the base tree and output, as the aggregation of shifted trees, the restriction on $[0;1)$ of

   $$f^m_{q,2^{-L}} \Big/ \int_{[0;1)} f^m_{q,2^{-L}}(v) dv.$$

Step 1a and Step 2 gather the ideas we have presented up until now, namely the definition and the aggregation of a finite number of TPT samples, with shifted underlying partitions. Another related distribution that we name the Continuous Pólya Aggregation (CPA) prior and denote CPA$(m, L, \mathcal{A}, U)$, corresponds to the situation where a continuum of Pólya Tree samples are aggregated with the operation (2.3). It is defined by a similar algorithm to the one above, except that Step 2 is replaced by

2'. Aggregation
Set

$$f = \sum_{i \in \mathbb{Z}} u_i H_{L,i}$$

as the base tree and output the restriction of $f^m_{\infty,2^{-L}}$ on $[0;1)$ as the aggregation of shifted trees.

Note that, according to Lemma 7, it is no longer necessary to normalize the output function as the u_i are defined so that $f^m_{\infty,2^{-L}}$ is almost surely a probability density. Since, as mentioned before, the external sets of shifted partitions are merged (see Figure 2.3), a 'Pólya forest' based on Steps 1a and 2 (or 2'.) only would have the side effect of cycling over the frontier (i.e., it tends the same limits toward the frontier), as shown in Figure 2.6. It presents a sample from a 'naive' construction, aggregating shifted trees without treatment on the frontier. That is why the above definitions of the DPA and the CPA prior also feature Steps 1b and 1c to modify the frontier behavior. It is possible to appreciate the advantage of this treatment on the frontier in Figures 2.7 and 2.8, where similar samples are plotted, with and without this modification. Without it, the samples would not be flexible enough to approximate general densities at a reasonable rate. The figures we have just mentioned also help understand the role of each parameter in the definition of DPA and CPA. The L parameter controls the depth of the trees (i.e., how refined the underlying partitions are), and, along with the degree of aggregation m, they define the smoothness of the prior. The number of trees q controls the distance between the samples from DPA and CPA. Indeed, the latter acts as a 'limit' prior for the former, and we most naturally found our theoretical analysis on it. As we will see, properties satisfied by CPA are shared with DPA, up to some discretization effects to be controlled. As one can observe on the figures, samples from DPA are piecewise constant so that it is still a histogram prior. However, in Section 2.5, it is shown that these samples are discrete approximations of spline functions, which are themselves sampled by CPA, and this added structure accounts for increased posterior performance (see Section 2.3).

Rather useful for adaptive estimation, U is a technical parameter related to our method to modify the samples on the frontier. A small value of U allows the samples to have limited complexity. This point is a prerequisite for our approach to derive posterior contraction rate. It is also the reason for the simplifications highlighted above, with the partitions' shifts being deterministic and the trees all derived from a single TPT sample.

2.3 Main results.

2.3.1 Posterior contraction rates for DPA.

We now present our main results on the asymptotic behavior of our density estimation procedure based on the DPA prior. In the following Theorem 2, we see that for any fixed arbitrary degree of Hölder regularity of the true density f_0, the posterior distribution attains minimax contraction rates (up to a logarithmic factor). A critical parameter that needs to be adequately defined to obtain the right degree of smoothness is the trees' depth L. In the following theorem setting, when the regularity of the true density f_0 is given, we let the depth depend on the sample size and α. Namely, we use the depth L_n that is the closest integer to the solution x of

$$2^x = \left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}}.$$

Theorem 2. *Suppose $f_0 \in \Sigma(\alpha, [0,1))$, $\alpha > 0$ and $f_0 \geq \rho$ for some $\rho > 0$. Let us endow f with a $DPA\left(\lfloor \alpha \rfloor, L_n, n, (a_l)_{0<l\leq L_n}, U_n\right)$ prior which we write Π, where $U_n \to \infty$ is an arbitrary sequence and such that, for some $\beta > 0, R \geq 1, \delta > 0$, for any $0 < l \leq L_n$,*

$$a_l \in \left[\delta n^{-\beta}; R\right].$$

Then, for $M > 0$ depending on ρ, α, $\|f_0\|_{\Sigma(\alpha)}$, β and R, and $d(f,g) = \|f - g\|_1$ or $d(f,g) = h(f,g)$, as $n \to \infty$,

$$\mathbb{E}_{f_0}\Pi\left[d(f_0, f) > M\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}} \mid X\right] \to 0.$$

Through the use of L_n and $\lfloor \alpha \rfloor$, this result relies on the assumption that the regularity α is fixed and known. The issue is that, in practice, we do not know this characteristic of the problem beforehand. Adding a prior on the depth of the trees L (which will be linked by the below functional to the degree of aggregation used) allows a new hierarchical prior attaining the minimax contraction rate, with no requirement for the knowledge of any fine property of f_0 anymore. We introduce the following functional, defined for $l, n \in \mathbb{N}^*$,

$$\xi(l,n) = \left\lfloor \frac{1}{2}\left[\frac{1}{l}\log_2\left(\frac{n}{\log n}\right) - 1\right]\right\rfloor.$$

For any depth value and sample size, it gives an estimate of the smoothness of the signal to be recovered. The result below shows that it is possible to define an adaptive prior that leads to optimal contraction rates for an arbitrary regularity $\alpha > 0$. The idea is to add a hyperprior on different models, characterized by the depth l and the corresponding estimated smoothness.

Theorem 3. *Suppose $f_0 \in \Sigma(\alpha, [0,1))$, $\alpha > 0$ and $f_0 \geq \rho$ for some $\rho > 0$. Let us endow f with the following hierarchical prior which we write Π,*

$$\begin{aligned} l &\sim \Pi_L \\ f \mid l &\sim DPA\left(\xi(l,n), l, v_n n \log^3 n, (a_l)_{0<l\leq L_n}, \log n\right), \end{aligned}$$

$v_n \to \infty$, and such that for $l > 0$, Π_L and the sequence $(a_l)_{l\in\mathbb{N}^*}$ satisfy, for some $\beta > 0, R \geq 1, \delta > 0$,

$$\Pi_L\left[\{l\}\right] \propto 2^{-l2^l} \qquad \text{and} \qquad a_l \in \left[\delta n^{-\beta}; R\right].$$

Then, for $M > 0$ depending on ρ, α, $\|f_0\|_{\Sigma(\alpha)}$, β and R, and $d(f,g) = \|f - g\|_1$ or $d(f,g) = h(f,g)$, as $n \to \infty$,

$$\mathbb{E}_{f_0}\Pi\left[d(f_0, f) > M\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}} \mid X\right] \to 0.$$

The exponential decay of the atom probabilities in Π_L is fast enough so that the prior still concentrates on a small number of models, but also slow enough so that it selects with high probability the model specified in Theorem 2 for a given α. From the point-of-view of Theorem 7 in Section 2.6, an essential tool to prove our theorems, this hyperprior is a good tradeoff

between the requirement that the prior concentrates its mass on a low-dimensional set and the one that it gives sufficient probability to small balls centered on the signal f_0. Therefore, our hierarchical prior behaves just like our non-adaptive one, attaining optimal contraction rates for any $\alpha > 0$.

A slight difference between Theorems 2 and 3 is that the sequence U_n is replaced by $\log n$. Indeed, as we seek to use Theorem 7 to prove the above result, the slow growth of the logarithmic function ensures that the adaptive prior concentrates its mass on sieves of moderate complexity. At the same time, neighbourhoods of f_0 still have sufficient mass asymptotically. Also, the number of trees is of a higher order since, in the adaptive setting, some further work is necessary to handle discretization effects of finite forests.

The novelty in these results is that these are the first tree-based priors that enjoy (almost-)optimal posterior contraction rates on classes of arbitrary smoothness to the best of our knowledge. It highlights how incorporating aggregation operations in priors leads to smoother forest samples when compared with tree priors. Previous results on related constructions were usually limited to Hölder classes of regularity $\alpha \leq 1$ at most. The link between single tree structures and piecewise constant functions makes them too rough to estimate smooth signals. In comparison with the original toy model of Arlot et al. [3], we extend the aggregation process so that the smoothing of the estimator occurs on regularity classes of orders even larger than 2. Also, we have shown that forest aggregation is compatible with the definition of adaptive priors, as these two aspects do not come at the price of a loss in posterior contraction rate. All in all, these are the first results of adaptivity and general smoothing for Bayesian forest estimators. To conclude on the advances made here, we mention that the original results on forest estimators from [3] put aside the effect of their construction on the frontier. The frequentist framework let them focus on a localized loss (namely, the mean integrated square error with integration on an interval strictly included in the unit interval), making the behavior on the frontier of $[0;1)$ irrelevant. On the contrary, the constructions we come up with here deal with those side effects by slightly modifying the samples near the frontier of Ω.

2.3.2 Extension to other priors.

Since the DPA prior is a discretized version of the CPA prior, the results from the previous section stem from the fact that it is possible to obtain similar ones for the CPA prior. Theorem 4 below is a version of Theorem 3 for CPA.

Theorem 4. *Suppose $f_0 \in \Sigma(\alpha, [0,1))$, $\alpha > 0$ and $f_0 \geq \rho$ for some $\rho > 0$. Let us endow f with the following hierarchical prior which we write Π*

$$\begin{aligned} l &\sim \Pi_L \\ f \,|\, l &\sim CPA(\xi(l,n), l, (a_i)_{0<i\leq l}, \log n) \end{aligned}$$

such that for $l > 0$, Π_L and the sequence $(a_l)_{l\in\mathbb{N}^*}$ satisfy, for some $\beta > 0, R \geq 1, \delta > 0$,

$$\Pi_L\left[\{l\}\right] \propto 2^{-l2^l} \qquad \text{and} \qquad a_l \in \left[\delta n^{-\beta}; R\right].$$

Then, for $M > 0$ depending on ρ, α, $\|f_0\|_{\Sigma(\alpha)}$, β and R, and $d(f,g) = \|f-g\|_1$ or $d(f,g) = h(f,g)$, as $n \to \infty$,

$$\mathbb{E}_{f_0}\Pi\left[d(f_0,f) > M\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}} \,|\, X\right] \to 0.$$

The CPA prior is a prior on spline densities which involves a randomized step to define the sample near the frontier of Ω. It is also possible to apply instead a deterministic correction to the infinite forest after the aggregation step. We now define the Spline Pólya Tree (SPT) prior which does so and further highlights the link between forests of Pólya Trees and spline priors.

Let's assume that $g \sim TPT_L(\mathcal{A})$ for some $L > 0$. Then, as g has support on Ω, we extend it on $\mathbb{R}$ by 1-periodicity, giving rise to the application $\tilde{g}$. For $m \geq 0$, let's define the map $A^1_{m,2^{-L}}$, which operates as a smoothing/aggregation of g, such that

$$A^1_{m,2^{-L}}(g) \coloneqq \tilde{g}^m_{\infty,2^{-L}}\Big|_{[0;1)}. \tag{2.6}$$

We now define the correction of this infinite forest. According to Lemma 8, since $\tilde{g}$ is a piecewise constant map with breaks at dyadic numbers $k2^{-L}$, $k \in \mathbb{Z}$, $A^1_{m,2^{-L}}(g)$ is a spline function of order $m+1$ and knots $\left(k2^{-L}\right)_{0\leq k\leq 2^L}$ (more about this in Section 2.5). Therefore, there exists polynomials $P_1,\ P_2$ of degree m such that, for $u \in \left[m2^{-L}; (m+1)2^{-L}\right)$ and $v \in \left[1-(m+1)2^{-L}; 1-m2^{-L}\right)$

$$A^1_{m2^{-L}}(g)(u) = P_1(u) \text{ and } A^1_{m,2^{-L}}(g)(v) = P_2(v).$$

As seen with CPA, the samples $A^1_{m,2^{-L}}(g)$ gives good estimates of a density on the interior of Ω but not near the frontier. An idea could be to modify the samples using only the information from $A^1_{m,2^{-L}}(g)$ away from the frontier, so that we then define the map

$$A^2_{m,2^{-L}}(g)(x) \coloneqq \begin{cases} A^1_{m,2^{-L}}(g)(x), & \text{if } x \in \left[m2^{-L}; 1-m2^{-L}\right), \\ P_1(x), & \text{if } x < m2^{-L}, \\ P_2(x), & \text{if } x \geq 1-m2^{-L}. \end{cases} \tag{2.7}$$

It is a spline function with no undesired continuity/cyclicality property. Finally, for $\tau > 0$, the following map is a density function

$$\tilde{f} = \frac{A^2_{m,2^{-L}}(g)_+ + \tau}{\int_0^1 \left(A^2_{m,2^{-L}}(g)_+(t) + \tau\right) dt}. \tag{2.8}$$

To sum up the above construction, if we write $SD_{\tau,m,2^{-L}}$ the application that associates the function $\tilde{f}$ to g, our SPT prior written $SPT(m, L, \mathcal{A}, \tau)$, is the image prior of a $TPT_L(\mathcal{A})$ prior by this map. Once again, this construction leads to adaptive (almost-)optimal contraction rate, for any arbitrary Hölder regularity, as is shown in the following theorem. However, the definition of samples near the frontier are less flexible than in CPA, which leads to an additional $\log$ factor in the rate. Still, it is not clear whether it is a shortcoming of SPT, or simply a byproduct of our proof.

Theorem 5. *Under the same assumptions on f_0 as in Theorem 4, for $\tau_n = \sqrt{n}^{-1}$, let's endow f with the following hierarchical prior which we write Π*

$$\begin{aligned} l &\sim \Pi_L \\ f \,|\, l &\sim SPT(\xi(l,n), l, \mathcal{A}, \tau_n) \end{aligned}$$

such that, for $l > 0$, Π_L and the sequence $(a_l)_{l\in\mathbb{N}^*}$ satisfy, for some $\beta > 0, R \geq 1, \delta > 0$,

$$\Pi_L[\{l\}] \propto 2^{-l^{3/2}2^l} \qquad \text{and} \qquad a_l \in \left[\delta n^{-\beta}; R\right].$$

Then, for $M > 0$ depending on ρ, α, $\|f_0\|_{\Sigma(\alpha)}$, β and R large enough and $d(f,g) = \|f-g\|_1$ or $d(f,g) = h(f,g)$, as $n \to \infty$,

$$\mathbb{E}_{f_0}\Pi\left[d(f_0,f) > Mn^{-\frac{\alpha}{2\alpha+1}}\log^{\frac{\alpha}{2\alpha+1}+1/2} n \,|\, X\right] \to 0.$$

2.4 Discussion.

One main take-away message is that a well-chosen histogram prior (in the form of a forest) achieves adaptation to arbitrary regularities $\alpha > 0$. We have shown that aggregating in a suitable way single (truncated) Pólya trees to define a forest prior allows the induced Hellinger and L^1 posterior contraction rates to be optimal on Hölder balls of densities of arbitrary smoothness coefficient. It then bypasses the apparent limitation to $\alpha \leq 1$ of single tree-based priors in previous works. This result highlights the benefits that aggregation operations have for Bayesian estimators. This also improves on previous results in the literature on forests estimators that assumed either $\alpha \leq 2$ or α fixed.

This work is a new step in the understanding of the theoretical behavior of Bayesian forest estimators. As noted above, it is still a "toy" model as it involves some simplification, in comparison with usual forest methods such as BART. Here, despite its definition as a sum-of-tree prior, the different trees in the forest are almost the same Pólya tree sample, with the difference that their underlying partitions are deterministically shifted. Whether it is possible to obtain results similar to those in this document with aggregation schemes and Bayesian forests that are more general is a matter for further investigation. A first extension would be to allow the shifts to be random. Or, one could allow the tree components to be defined on deterministic partitions of the sample space but with different values in corresponding cells across the trees. Our results seem to pertain more particularly to priors on forests of many well-correlated trees, and it is natural to try to lighten this imposed correlation. Doing so, we would obtain priors closer to those used in practice.

Here we focused on the density model, but the ideas developed should more generally apply to other settings, e.g., nonparametric regression, which is left for further work. The framework of Theorem 7 in Section 2.6 limits our analysis of the Bayesian density estimation to Hellinger and L^1 rates. Other distances require more complex arguments involving additional technicalities. However, we mention that this framework extends to the L^2 distance in regression models. Then, we expect our construction and the rates we obtained to apply to these models as well.

Another intriguing research direction is extending this work to higher dimensions, where the sample space is instead $[0;1)^d$, with d potentially large. Though Pólya Trees in this setting exist, the definition of forest aggregations is not straightforward. It is necessary to define the partition shifts carefully so that the 'limiting continuous' prior on 'continuous forests' enjoys good properties. We see in the next section that the elements of a B-spline basis appear naturally in the limit with our aggregation of piecewise constant maps. In higher dimensions, with well-defined shifts, it should be possible to recreate a tensor-product basis so that our analysis applies as well.

Also, we have left aside the question of the computation of the posterior induced by our new priors. It was not the primary purpose of this article, and consequently, we do not investigate this further here. Even though the conjugacy property of single Pólya Trees is lost through aggregation, usual methods such as MCMC should apply. Although we discussed its

shortcomings, a simplified prior with no treatment near the frontier of Ω has the convenient upside that its posterior is explicit. To illustrate the behavior of such a Bayesian forest, we present some numerical experiments based on this prior in Section 2.6.2.

Finally, in the present work, we adapt ideas developed in the study of frequentist estimators in [3] to obtain optimal Bayesian posterior rates for arbitrary regularities. Yet, extending these new optimality results to frequentist random forests is not straightforward. Interestingly, the Bayesian framework seems more conducive to developing methods that attain optimal rates for high regularities: for instance, Bayesian mixtures of Gaussians can adapt to arbitrary regularities (see [92]), even though the Gaussian kernel has order 1. This contrasts with frequentist Gaussian kernel estimators, which are suboptimal for higher regularities. Therefore, Bayesian forest posteriors could also have an advantage over frequentist random forests.

2.5 Proofs.

2.5.1 Link with spline spaces.

First, for A a real interval, we denote $\Pi_{k,\mathbf{t}}(A)$ the space of splines of order k and knot sequence $\mathbf{t} = (t_i)_{i\in I}$, with $I \subset \mathbb{Z}$ such that $\forall i,\ t_i \in A \subset \mathbb{R}$. Also, we assume $i \leq j \implies t_i \leq t_j$ as well as, $\inf\{t_i \,|\, i \in I\} = \inf A$ and $\sup\{t_i \,|\, i \in I\} = \sup A$. $\Pi_{k,\mathbf{t}}$ is the subset of maps in $\mathcal{C}^{k-2}(A)$ whose restriction on any interval of the form $[t_i; t_{i+1}[$ is a polynomial of degree strictly smaller than k. It coincides with the span of B-splines $B_{j,k}$ of order k and knots $\mathbf{t}$ (see [46] or [65] for definition and more details on this), with $j = 1, \ldots, k + \#\mathbf{t} - 1$ if $\#\mathbf{t} < \infty$, $j \in \mathbb{Z}$ otherwise (dropping the dependence on the sequence of t_j's as there will be no ambiguity in the following). Now, as shown in [5], for $L \in \mathbb{N}$, a real sequence $(u_i)_{i\in\mathbb{Z}}$ and the piecewise constant map

$$\begin{aligned} \widetilde{H}_L \colon \mathbb{R} &\to \mathbb{R} \\ x &\mapsto \sum_{j\in\mathbb{Z}} u_j H_{Lj}, \end{aligned} \tag{2.9}$$

we write for $m \in \mathbb{N}$, by linearity and from (2.4),

$$\begin{aligned} \left(\widetilde{H}_L\right)^m_{\infty,2^{-L}}(x) &= \sum_{j\in\mathbb{Z}} 2^L u_j \chi^{*m}(2^L\cdot) * H_{Lj}(x) \\ &= \sum_{j\in\mathbb{Z}} u_j 2^{2L} \chi^{*m}(2^L\cdot) * \chi(2^L\cdot - j)(x) \\ &= \sum_{j\in\mathbb{Z}} u_j 2^{2L} \int_{\mathbb{R}} \chi^{*m}(2^L s)\chi\left(2^L(x-s) - j\right) ds \\ &= \sum_{j\in\mathbb{Z}} u_j 2^L \chi^{*(m+1)}\left(2^L x - j\right). \end{aligned} \tag{2.10}$$

Also, as shown in [46] (Section 10), the maps $\{\chi^{*(m+1)}(\cdot - i), i \in \mathbb{Z}\}$ are the Cardinal splines of order $m+1$, i.e., the B-Splines of order $m+1$ corresponding to the biinfinite knot sequence $\mathbf{t} = \mathbb{Z}$. Hence, it is a basis for $\Pi_{m+1,\mathbb{Z}}(\mathbb{R})$ and similarly, one shows that $\{2^L\chi^{*(m+1)}(2^L\cdot - i) \,|\, i \in \mathbb{Z}\}$ is a basis for $\Pi_{m+1,2^{-L}\mathbb{Z}}(\mathbb{R})$. Therefore, $(\widetilde{H}_L)^m_{\infty,2^{-L}}$ from (2.10) belongs to this linear space and $(u_i)_{i\in\mathbb{Z}}$ is the sequence of its coordinates in this basis.

We now remark that the map from (2.10) is 1-periodic if and only if $(u_i)_{i\in\mathbb{Z}}$ is a 2^L-periodic sequence (see Lemma 6). It follows that $\tilde{\Pi}_{m+1,2^{-L}\mathbb{Z}}(\mathbb{R}) \coloneqq \Pi_{m+1,2^{-L}\mathbb{Z}}(\mathbb{R}) \cap \{f : \mathbb{R} \to \mathbb{R},\ f\ 1-\text{periodic}\}$ is a linear space with basis

$$\left\{ S_{i,2^L,m} \coloneqq 2^L \sum_{j\in\mathbb{Z}} \chi^{*(m+1)}(2^L \cdot -2^L j - i) \,|\, 0 \le i \le 2^L - 1 \right\}. \tag{2.11}$$

In Section 2.5.3, we see that this space has good approximation properties, which we use below in the proofs. Indeed, for each Hölderian density function on Ω, it contains an element whose restriction to Ω is sufficiently close to the density function in the interior of the interval. We also stress that, via a modification near the frontier of Ω, we can recover a spline function that approximates the density well on the whole interval.

This observation accounts for the performance of our priors. Looking back at the algorithmic definitions of DPA and CPA priors in Section 2.2.3, and setting aside the discretization in DPA, it is possible to interpret them as follows. The coordinates of an element $\tilde{\Pi}_{m+1,2^{-L}\mathbb{Z}}(\mathbb{R})$ are sampled via a TPT distribution, and uniform random variables transform it into a similar spline from $\Pi_{m+1,2^{-L}\mathbb{Z}}(\mathbb{R})$ (or rather a restriction to Ω). This point of view underlies the proofs in the following subsections.

2.5.2 Proofs of main results.

In this subsection, we provide the proofs for the theorems presented in the last section. The adaptive results involve the derivation of intermediary points that we first demonstrate in the proof of non-adaptive results (for fixed regularity). Therefore, we first analyze the case of fixed regularities for CPA and DPA before delving into the proofs of adaptive results. Our arguments rely on Theorem 7, whose conditions are investigated in lemmas following it. These lemmas build on the approximation properties of spline functions and their parametric representations (see Section 2.5.3). Also, we primarily focus on CPA, as the similar properties of DPA only then require the control of additional discretization terms.

Proof of Theorem 2 (and extension to CPA for fixed regularity). It is sufficient to verify that the conditions of Theorem 7 are satisfied. One shows that the prior puts sufficient mass in some Kullback-Leibler neighborhoods of the true density. We use results in Approximation Theory (see Lemmas 1 and 10) that we develop in Sections 2.5.3 and 2.6. Besides, one also has to prove that the priors allocate most of their mass to subsets of limited complexity. It ensues from the priors generating draws that belong to spaces that resemble spaces of splines, whose dimensions are not too large (see Section 2.5.1 and Lemma 11). In the following, we write $m = \lfloor \alpha \rfloor$ and, for c_0 a constant to be defined below,

$$\epsilon_n = c_0 \left(\frac{\log n}{n} \right)^{\frac{\alpha}{2\alpha+1}}. \tag{2.12}$$

1) Complexity of the prior: Let us define

$$\mathcal{H}_{m,l} \coloneqq \left\{ (\theta_i)_{i\in\mathbb{Z}} \in \mathbb{R}_+^{\mathbb{Z}}, \quad \forall i \in \mathbb{Z},\ \theta_i = \theta_{i+m+2^l}, \right. \\ \left. \sum_{i=0}^{2^l-m-1} \theta_i + \sum_{i=2^l-m}^{2^l-1} \left(\omega_{m,2^l-i}\theta_i + \omega_{m,i-(2^l-m-1)}\theta_{i+m} \right) = 1 \right\}. \tag{2.13}$$

According to Lemma 7 and the discussion following it, the following sets have probability 1 under the DPA prior (resp. CPA):

$$\mathcal{F}_n := \left\{ \left. \frac{f^m_{n,2^{-L_n}}}{\int_0^1 f^m_{n,2^{-L_n}}(t)dt} \right|_{[0;1)}, \quad f = \sum_{i\in\mathbb{Z}} \theta_i H_{L_n i},\ (\theta_i)_{i\in\mathbb{Z}} \in \mathcal{H}_{m,L_n} \right\} \tag{2.14}$$
$$\left(\text{resp. } := \left\{ \left. f^m_{\infty,2^{-L_n}} \right|_{[0;1)}, \quad f = \sum_{i\in\mathbb{Z}} \theta_i H_{L_n i},\ (\theta_i)_{i\in\mathbb{Z}} \in \mathcal{H}_{m,L_n} \right\} \right).$$

In (2.5), the positivity of χ ensures that $\inf_{1\leq l\leq m+1} \omega_{m,l} = \omega_{m,1} = 1/(m+1)!$ (see Proposition 6.7.1., p.136, in [5]). So, any sequence in $\mathcal{H}_{m,n}$ has its coordinates bounded by $(m+1)!$ because of their positivity and the constraint from the definition. Then, from Lemma 11 with $M = (m+1)!$, $q = n$ and $L = L_n$, there exists an absolute constant C such that for n large enough and $B_{\mathbb{R}^D}(0,r)$ the L^2 closed ball in $\mathbb{R}^D$ of radius r,

$$\begin{aligned} N(\epsilon_n, \mathcal{F}_n, h) &\leq N\left(C\left(2^{L_n}+m\right)^{-1/2} \epsilon_n^2, \mathcal{H}_{m,n}, \|\cdot\|_2 \right) \\ &\leq N\left(C\left(2^{L_n}+m\right)^{-1/2} \epsilon_n^2, B_{\mathbb{R}^{2^{L_n}+m}}\left(0, \sqrt{2^{L_n}+m}\omega_{m,1}^{-1}\right), \|\cdot\|_2 \right). \end{aligned}$$

The first inequality is valid in the discrete case since the remainder term from Lemma 11 is of the order $o(\epsilon_n)$ for our values of L_n and the number of trees $q = n$. It is a general fact that there exists a universal constant $C > 0$ such that

$$N(\delta, B_{\mathbb{R}^K}(0,M), ||.||_2) \leq \left(\frac{CM}{\delta}\right)^K.$$

Therefore, one concludes that, for any $D > c_0^{-1}$ and n large enough

$$N(\epsilon_n, \mathcal{F}_n, h) \leq \left(C\omega_{m,1}^{-1} \frac{2^{L_n}+m}{\epsilon_n^2} \right)^{2^{L_n}+m} \leq e^{Dn\epsilon_n^2}. \tag{2.15}$$

2) Small ball probability condition: For the last condition of Theorem 7, we introduce the sequence $(\eta_i)_{0\leq i\leq 2^{L_n}+m-1}$ from Lemma 1 such that

$$1 = \sum_{i=m}^{2^{L_n}-1} \eta_i + \sum_{i=0}^{m-1} \left(\omega_{m,i+1}\eta_i + (1-\omega_{m,i+1})\eta_{2^{L_n}+i}\right).$$

We define $\tilde{\eta}_i = \eta_i$ for $i = m, \ldots, 2^{L_n}-1$ and for $i = 0, \ldots, m-1$,

$$\tilde{\eta}_i = \tilde{\eta}_{2^{L_n}+i} = \omega_{m,i+1}\eta_i + (1-\omega_{m,i+1})\eta_{2^{L_n}+i} = (1-\omega_{m,m-i})\eta_i + \omega_{m,m-i}\eta_{2^{L_n}+i}, \tag{2.16}$$

this being consistent according to Lemma 5. This guarantees that $(\tilde{\eta}_i)_{0\leq i\leq 2^{L_n}-1} \in S^{2^{L_n}}$. First, from Lemma 10, for n and c_0 large enough and C small enough, depending on ρ, α and $\|f_0\|_{\Sigma(\alpha)}$, we have the inequality

$$\Pi\left[B_{KL}\left(f_0, \epsilon_n\right) \right] \geq \Pi\left[\max_{0\leq i\leq 2^{L_n}+m-1} |u_{i-m} - \eta_i| \leq C\epsilon_n 2^{-L_n} \right]$$

following from the fact that the terms depending on $L = L_n$ and the number of 'trees' $q = n$ in Lemma 10 are of order $O(\epsilon_n^2)$. One controls the different random variables from the

sequential definition of our prior in Step 1 so that one obtains a lower bound on the above event probability. With notation from Part 2.2.3 and $\iota(i) \equiv i + m \mod 2^{L_n}$, for any $r > 0$, on the event

$$\left\{ \max_{0 \leq i \leq 2^{L_n}-1} \left| \Theta_i - \tilde{\eta}_{\iota(i)} \right| \leq \frac{\omega_{m,1}^2}{8} r \leq \frac{\omega_{m,1}}{8} r \quad \text{and} \quad \max_{2^{L_n}-m \leq i \leq 2^{L_n}-1} |\theta_i - \eta_{i+m}| \leq \omega_{m,1} \frac{r}{4} \right\},$$

we have that $\max_{m \leq i \leq 2^{L_n}+m-1} |u_{i-m} - \eta_i| \leq r$ and, using the periodicity of $(u_i)_{i \in \mathbb{Z}}$, for $i = 0, \dots, m-1$,

$$\begin{aligned}
|u_{i-m} - \eta_i| &= |u_{i+2^{L_n}} - \eta_i| \\
&= |\theta_{i+2^{L_n}} - \eta_i| \\
&= \left| \frac{\Theta_{i+2^{L_n}-m} - \omega_{m,m-i}\theta_{i+2^{L_n}-m}}{1-\omega_{m,m-i}} - \frac{\tilde{\eta}_i - \omega_{m,m-i}\eta_{2^{L_n}+i}}{1-\omega_{m,m-i}} \right| \\
&\leq (1-\omega_{m,m-i})^{-1} \left(\frac{\omega_{m,1}}{8} r + \omega_{m,m-i}\omega_{m,1} \frac{r}{4} \right) \\
&\leq r(1/8 + \omega_{m,m-i}/4) \leq r.
\end{aligned}$$

This ultimately implies that $\max_{0 \leq i \leq 2^{L_n}+m-1} |u_{i-m} - \eta_i| \leq r$ since $\omega_{m,1} \leq \omega_{m,l} \leq 1$ for $l \geq 1$. Therefore, it remains to study the factors in the lower bound

$$\begin{aligned}
\Pi \left[B_{KL}\left(f_0, \epsilon_n \right) \right] \geq \Pi \left[\max_{0 \leq i \leq 2^{L_n}-1} \left| \Theta_i - \tilde{\eta}_{\iota(i)} \right| \leq C \frac{\omega_{m,1}^2}{8} \epsilon_n 2^{-L_n} \right] \times \\
\Pi \left[\max_{2^{L_n}-m \leq i \leq 2^{L_n}-1} |\theta_i - \eta_{i+m}| \leq C\omega_{m,1} \frac{\epsilon_n}{4} 2^{-L_n} \middle| \max_{0 \leq i \leq 2^{L_n}-1} \left| \Theta_i - \tilde{\eta}_{\iota(i)} \right| \leq C \frac{\omega_{m,1}^2}{8} \epsilon_n 2^{-L_n} \right].
\end{aligned}$$

This translates the fact that, to obtain a good prior mass, it is sufficient to control the mass of the TPT so that the associated forests are close to the density f_0 on the interior of Ω, and then control the behavior near the frontier to extend the result to the whole of Ω.

Now, for $0 \leq i \leq 2^{L_n} - 1$, we decompose $\Theta_i = \prod_{j=1}^{L_n} Y_{\kappa(L_n,i)[j]}$ and $\tilde{\eta}_i = \prod_{j=1}^{L_n} y_{\kappa(L_n,i)[j]}$ where, for $1 \leq j < L_n$,

$$y_{\kappa(L_n,i)[j]0} := \frac{\sum_{s, \lfloor s 2^{j+1-L_n} \rfloor_f = \lfloor i 2^{j+1-L_n} \rfloor_f} \tilde{\eta}_s}{\sum_{s, \lfloor s 2^{j-L_n} \rfloor_f = \lfloor i 2^{j-L_n} \rfloor_f} \tilde{\eta}_s}, \qquad y_{\kappa(L_n,i)[j]1} := 1 - y_{\kappa(L_n,i)[j]0}$$

belong to $[0; 1]$. Also, y_0 and y_1 satisfy the same formula, with $j = 0$. Withh $e_j = Y_{\kappa(L_n,i)[j]}$ and $t_j = y_{\kappa(L_n,\iota(i))[j]}$ for sake of clarity, then, as for all $j = 1, ..., L_n$, $|e_j| \leq 1$ and $|t_j| \leq 1$, we have

$$\begin{aligned}
\left| \Theta_i - \tilde{\eta}_{\iota(i)} \right| &= \left| \sum_{j=1}^{L_n} e_1 ... e_{j-1}(e_j - t_j) t_{j+1} ... t_{L_n} \right| \\
&\leq \sum_{j=1}^{L_n} |e_1 ... e_{j-1}(e_j - t_j) t_{j+1} ... t_{L_n}| \\
&\leq \sum_{j=1}^{L_n} |e_j - t_j| .
\end{aligned}$$

This finally gives us that, using that the Y's variables are independent in the TPT,

$$\Pi\left[\max_{0\le i\le 2^{L_n}-1}\left|\Theta_i-\tilde\eta_{\iota(i)}\right|\le\frac{C\omega_{m,1}^2\epsilon_n}{8*2^{L_n}}\right]\ge\Pi\left[\bigcap_{\substack{0\le i\le 2^{L_n},\\ 1\le j\le L_n}}\left\{\left|Y_{\kappa(L_n,i)^{[j]}}-y_{\kappa(L_n,\iota(i))^{[j]}}\right|\le\frac{C\omega_{m,1}\epsilon_n^2}{8L_n2^{L_n}}\right\}\right]$$
$$=\prod_{j=1}^{L_n}\prod_{|\kappa|=j-1}P_{X\sim\mathsf{Beta}(a_j,a_j)}\left[|X-y_{\kappa 0}|\le C\frac{\omega_{m,1}^2}{8L_n}\epsilon_n2^{-L_n}\right].$$

Let's write $\xi_n=C\frac{\omega_{m,1}^2}{8L_n}\epsilon_n2^{-L_n}$. Since for any j, $a_j\Gamma(a_j)=\Gamma(a_j+1)\le\Gamma(R+1)=:\tilde R$ and Γ is lower bounded by some constant $\psi>0$ on the set of real positive numbers, $\Gamma(2a_j)\Gamma(a_j)^{-2}\ge\psi a_j^2\tilde R^{-2}\ge\psi\tilde R^{-2}\delta^2n^{-2\beta}$. Also, for n large enough, if $R\ge a_j>1$,

$$\int_{(y_{\kappa 0}-\xi_n)\vee 0}^{(y_{\kappa 0}+\xi_n)\wedge 1}t^{a_j-1}(1-t)^{a_j-1}dt\ge\int_0^{\xi_n}t^{a_j-1}(1-t)^{a_j-1}dt\ge(1-\xi_n)^{R-1}\frac{\xi_n^R}{R}\ge\frac{\xi_n^R}{2^{R-1}R},$$

while, for $a_j\le 1$, the bound can just be replaced by ξ_n when n is large enough. Finally, for some $C>0$, depending on β, R, and c_0,

$$\begin{aligned}\Pi\left[\max_{0\le i\le 2^{L_n}-1}\left|\Theta_i-\tilde\eta_{\iota(i)}\right|\le\tfrac{C\omega_{m,1}^2\epsilon_n}{8*2^{L_n}}\right]&\ge\Pi_{j=1}^{L_n}\Pi_{|\kappa|=j-1}\tfrac{\psi\delta^2}{\tilde R^2}\left(\tfrac{1}{2^{R-1}R}\wedge 1\right)n^{-2\beta}\xi_n^{R\vee 1}\\&=\left(\tfrac{\psi\delta^2}{\tilde R^2}\left(\tfrac{1}{2^{R-1}R}\wedge 1\right)n^{-2\beta}\xi_n^{R\vee 1}\right)^{2^{L_n}-1}\\&\ge e^{-Cn\epsilon_n^2}.\end{aligned}\tag{2.17}$$

Then, for U_n large enough, i.e. n large enough, for any $2^{L_n}-m\le i\le 2^{L_n}-1$, the uniform random variables θ_i verify for any $\tilde C>0$

$$\begin{aligned}\Pi\left[|\theta_i-\eta_{i+m}|\le C\omega_{m,1}\tfrac{\epsilon_n}{4}2^{-L_n}\;\middle|\;\max_{0\le i\le 2^{L_n}-1}\left|\Theta_i-\tilde\eta_{\iota(i)}\right|\le C\tfrac{\omega_{m,1}^2}{8}\epsilon_n2^{-L_n}\right]&\\\ge\left(\tfrac{\Theta_i}{\omega_{m,\iota(i)}}\right)^{-1}C\omega_{m,1}\tfrac{\epsilon_n}{8}2^{-L_n}&\\\ge C\omega_{m,1}^2\epsilon_n2^{-L_n}/8\ge e^{-\tilde Cn\epsilon_n^2}.&\end{aligned}$$

The first inequality is due to $0\le\eta_{i+m}\le\omega_{m,\iota(i)+1}^{-1}\tilde\eta_{\iota(i)}\le\omega_{m,\iota(i)+1}^{-1}\left(\Theta_i+C\omega_{m,1}^2\epsilon_n2^{-L_n}/8\right)$ on the conditioning event, which follows by positivity and (2.16). Finally, we use the conditional independence of the random variables θ_i to obtain the lower bound,

$$\Pi\left[B_{KL}\left(f_0,\epsilon_n\right)\right]\ge e^{-(C+m\tilde C)n\epsilon_n^2}.\tag{2.18}$$

We now conclude with Theorem 7 and equations (2.15) and (2.18), recalling that $\mathcal{F}_n$ is an almost sure event under our prior, the constant $M>0$ depending on ρ,α, $\|f_0\|_{\Sigma(\alpha)}$, β and R. □

Proof of Theorems 3 and 4. We proceed as in the proof of Theorem 2, with ϵ_n as in (2.12). The main difference is that the distributions now allocate positive mass to different depth values L so that we adapt the sieves. Below, we take the union of sieves similar to those introduced in the above proof, from low resolution $L=1$ up to some threshold. To this effect, we introduce the sequences of depth $L_{1,n}$ and $L_{2,n}$ such that $2^{L_{i,n}}\asymp C_i\left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}}$ (i.e., it is

the closest integer to the solution of this equation) for some constants C_1 and $C_2 = 1$. The proof then again uses the Theorem 7 with an additional term to be controlled, corresponding to the prior mass on the hyperparameter L.

1) Complexity of the prior: Let, for $0 \leq k \leq 2^{L_n} - 1$,

$$\mathcal{I}_{k,n} := \Bigg\{ (\theta_i)_{i\in\mathbb{Z}} \in \mathbb{R}_+^{\mathbb{Z}},\ \forall i \in \mathbb{Z},\ \theta_i = \theta_{i+k+2^{L_{1,n}}} \text{ and } 0 \leq \theta_i \leq \log n, \\ \sum_{i=0}^{2^{L_{1,n}}-k-1} \theta_i + \sum_{i=2^{L_{1,n}}-k}^{2^{L_{1,n}}-1} \left(\omega_{k,2^{L_{1,n}}-i}\theta_i + \omega_{k,i-(2^{L_{1,n}}-k-1)}\theta_{i+k} \right) = 1 \Bigg\}$$

and define the sieves for the DPA prior (resp. CPA)

$$\mathcal{F}_n := \bigcup_{l=1}^{L_{1,n}} \left\{ f_{\infty,2^{-l}}^{\xi(l,n)}[0;1),\ f = \sum_{i\in\mathbb{Z}} \theta_i H_{li},\ (\theta_i)_{i\in\mathbb{Z}} \in \mathcal{I}_{\xi(l,n),n} \right\} \tag{2.19}$$

$$\left(\text{resp. } := \bigcup_{l=1}^{L_{1,n}} \left\{ \frac{f_{n\log^3 n,2^{-l}}^{\xi(l,n)}}{\int_0^1 f_{n\log^3 n,2^{-l}}^{\xi(l,n)}(t)dt} \Bigg|_{[0;1)},\ f = \sum_{i\in\mathbb{Z}} \theta_i H_{li},\ (\theta_i)_{i\in\mathbb{Z}} \in \mathcal{I}_{\xi(l,n),n} \right\} \right).$$

In the definition of the prior, the sequence $(u_i)_{i\in\mathbb{Z}}$ lies in $[0;\log n]^{\mathbb{Z}}$ almost surely for n large enough. Therefore, following Lemma 7 and the discussion after its proof, we now have, for n large enough,

$$\begin{aligned} \Pi[\mathcal{F}_n^c] = \Pi[l > L_{1,n}] \propto \sum_{l=L_{1,n}+1}^{+\infty} 2^{-l2^l} &\lesssim 2^{-L_{1,n}2^{L_{1,n}}} \\ &= e^{-\log(C_1 n/\log n)\frac{2^{L_{1,n}}}{2\alpha+1}} \\ &\leq e^{-C_1(c_0^2/(4\alpha-2))^{-1} n\epsilon_n^2}. \end{aligned} \tag{2.20}$$

Also, using Lemma 11 with $M = \log n$, $q = v_n n\log^3 n$, $m \leq \xi(1,n) \leq \log(n)/2$ and $l \leq L_{1,n}$, we use similar arguments as the ones preceding (2.15) to derive, for C, C' absolute constants and D depending on C_1 and c_0,

$$\begin{aligned} N(\epsilon_n, \mathcal{F}_n, h) &\leq \sum_{l=1}^{L_{1,n}} \left(\frac{C(2^l + \xi(l,n))\log n}{\epsilon_n^2} \right)^{2^l+\xi(l,n)} \\ &\leq \sum_{l=1}^{L_{1,n}} \left(\frac{C(2^{L_{1,n}} + \log(n)/2)\log n}{\epsilon_n^2} \right)^{2^l+\log(n)/2} \\ &\lesssim \left(\frac{C(2^{L_{1,n}} + \log(n)/2)\log n}{\epsilon_n^2} \right)^{2^{L_{1,n}}+\log n} \leq e^{C'\log n 2^{L_{1,n}}} \leq e^{Dn\epsilon_n^2}. \end{aligned} \tag{2.21}$$

In particular, we have used that, with the sequences from the theorem, the term depending on q in Lemma 11 is of order $o(\epsilon_n)$.

2) Prior mass condition: Since $\xi(L_{2,n}, n) = \lfloor \alpha \rfloor$, it is possible to use the same arguments that led to (2.18), for n large enough, c_0 large enough depending on ρ, α and $\|f_0\|_{\Sigma(\alpha)}$ and C large enough depending on β, R and c_0, to obtain

$$\Pi\left[B_{KL}(f_0,\epsilon_n)\right] \gtrsim \Pi\left[B_{KL}(f_0,\epsilon_n)\,|\,l=L_{2,n}\right] 2^{-L_{2,n}2^{L_{2,n}}} \geq e^{-Cn\epsilon_n^2} e^{-(c_0^2/(2\alpha-1))^{-1}n\epsilon_n^2}. \tag{2.22}$$

Indeed, in the argument invoking Lemma 10, the terms controlled with $q = v_n n \log^3 n$ are of order $o(\epsilon_n^2)$.

We conclude using Theorem 7 along with equations (2.20),(2.21) and (2.22), since for C_1 large enough, $C_1(c_0^2/(4\alpha-2))^{-1} > C + (c_0^2/(2\alpha-1))^{-1} + 4$. Then, the theorem is valid for M large enough, depending on ρ, α, $\|f_0\|_{\Sigma(\alpha)}$, β and R. □

2.5.3 Approximation theory for periodic splines.

In the constructions of CPA and DPA distributions, the aggregating operation we have defined transforms a TPT sample into a periodic spline density (or a piecewise constant approximation of it for DPA). It is convenient as these periodic splines have good approximation properties according to the following lemma. It is the result of (2.10), Lemmas 6 and 8, and the prior definitions. Below, we prove that it is possible to approximate any Hölder density with such a spline, as long as we focus on an interval far enough from the frontier of Ω.

In order to extend this result near the frontier of Ω, we also see that we can recover an approximating spline on the whole of Ω from the periodic spline of Lemma 1. Consequently, CPA and DPA include a stochastic step to simulate the modifications needed to obtain this last spline density.

In the end, the link with splines explains why the aggregation part of our priors results in the almost optimal contraction rates that we obtained in Section 2.3.

Lemma 1. *Suppose $m+1 \geq \alpha > 0$ and $L \geq 1$. There exists a constant C depending only on m and α such that for every $f_0 \in \Sigma(\alpha,[0,1))$, there exists $g \in \tilde{\Pi}_{m+1,2^{-L}\mathbb{Z}}(\mathbb{R})$ such that*

$$\left\| f_0\big|_{[2^{-L}m;1-2^{-L}m)} - g\big|_{[2^{-L}m;1-2^{-L}m)} \right\|_\infty \leq C2^{-\alpha L}\left(\left\|f_0^{(\lfloor\alpha\rfloor)}\right\|_\infty + \|f_0\|_{\Sigma(\alpha)}\right)$$

for L large enough.
Let f_0 be a probability density such that $f_0 > \rho$ for some $\rho > 0$. For L large enough, replacing the above bound by $C2^{-\alpha L}$ with C a constant depending on m, α, $\left\|f_0^{(\lfloor\alpha\rfloor)}\right\|_\infty$ and $\|f_0\|_{\Sigma(\alpha)}$, we can choose g above of the form, for $S_{i,2^L,m}$ as in (2.11),

$$g = \sum_{i=0}^{2^L-1} \theta_i S_{i,2^L,m}$$

with $\theta = (\theta_i)_{0\leq i\leq 2^L}$ in the 2^L-dimensional unit simplex S^{2^L}, such that there exists

$$(\eta_i)_{0\leq i\leq 2^{L}+m-1} \in \left[0; 2\Big(\|f_0\|_{\Sigma(\alpha)} + 2\left\|f_0^{(\lfloor\alpha\rfloor)}\right\|_\infty\Big)\right]^{2^L+m}$$

satisfying

$$\theta_k = \begin{cases} \eta_k & \text{if } m \leq k \leq 2^L-1 \\ \omega_{m,k+1}\eta_k + (1-\omega_{m,m-k})\eta_{2^L+k} & \text{if } 0 \leq k \leq m-1 \end{cases}$$

and

$$\left\| f_0 - \sum_{k=0}^{2^L+m-1} \eta_k 2^L \chi^{*(m+1)}(2^L \cdot -(k-m)) \Big|_{[0;1)} \right\|_\infty \leq C 2^{-\alpha L}.$$

Proof. Let's introduce the B-spline functions of order $m+1$ on the interval $[-m2^{-L}; 1+m2^{-L}]$ corresponding to the knots $i2^{-L},\ -m \leq i \leq 2^L+m$, denoted $B_{1,m+1}, \ldots, B_{2^L+3m,m+1}$. Figure 2.4 depicts these basis functions in the particular case $L=3$ and $m=3$.

The Cox-de Boor recursion formula ensures that B-splines whose supports are far enough from the edges $-m2^{-L}$ and $1+m2^{-L}$ are actually Cardinal splines with suitable scaling. As shown in [46] (Section 10),

$$B_{k,m+1} = \chi^{*(m+1)}\left(2^L \cdot -(k-2m-1)\right), \qquad m+1 \leq k \leq 2^L+2m. \tag{2.23}$$

Also, $B_{i,m+1}$ is supported in an interval of length at most $(m+1)2^{-L}$ included in $[(i-(2m+1))2^{-L}; (i-m)2^{-L}]$, i.e.,

$$\forall x \notin [(i-(2m+1))2^{-L}; (i-m)2^{-L}], \qquad B_{i,m+1}(x) = 0. \tag{2.24}$$

As $f_0 \in \Sigma(\alpha, [0,1))$, according to Lemma 16, there exists a map

$$h:\ [-(m+1)2^{-L}; 1+(m+1)2^{-L}] \to \mathbb{R} \text{ such that,}$$
$$h \in \Sigma\left(\alpha, [-m2^{-L}; 1+m2^{-L}]\right),\ \|h\|_{\Sigma(\alpha)} = \|f_0\|_{\Sigma(\alpha)},\ h\Big|_{[0;1)} = f_0.$$

Also, for L large enough, $\left\|h^{(\lfloor\alpha\rfloor)}\right\|_\infty \leq 2\left\|f_0^{(\lfloor\alpha\rfloor)}\right\|_\infty$ by continuity. Using Lemma 9 and (2.24), there exists C depending only on m and α, and reals $\theta_k,\ m+1 \leq k \leq 2^L+2m$, bounded by $\|f_0\|_{\Sigma(\alpha)} + 2\left\|f_0^{(\lfloor\alpha\rfloor)}\right\|_\infty$, such that for L large enough

$$\begin{aligned}\left\| h\Big|_{[0;1)} - \sum_{k=m+1}^{2^L+2m} \theta_k B_{k,m+1} \Big|_{[0;1)} \right\|_\infty &\leq C 2^{-\alpha L}\left(\left\|h^{(\lfloor\alpha\rfloor)}\right\|_\infty + \|h\|_{\Sigma(\alpha)}\right) \\ &\leq C 2^{-\alpha L}\left(\left\|f_0^{(\lfloor\alpha\rfloor)}\right\|_\infty + \|f_0\|_{\Sigma(\alpha)}\right).\end{aligned} \tag{2.25}$$

In addition, thanks to the small support of $\chi^{*(m+1)}$ (see Lemma 2) and the equality (2.23), the maps $\sum_{k=m+1}^{2^L+2m} \theta_k B_{k,m+1}$ and, with $\tilde{k}(k) = (k-2m-1 \bmod 2^L)$,

$$\sum_{k=m+1}^{2^L+m} \theta_k \sum_{i\in\mathbb{Z}} \chi^{*(m+1)}\left(2^L \cdot -(k+i2^L-2m-1)\right) = \sum_{k=m+1}^{2^L+m} \theta_k 2^{-L} S_{\tilde{k}(k),2^L,m}(\cdot)$$

are equal on the interval $\left[2^{-L}m; 1-2^{-L}m\right)$. The latter map then satisfies the inequality in the first part of the theorem according to (2.25) and belongs to $\tilde{\Pi}_{m+1,\mathbb{Z}/q}(\mathbb{R})$ following (2.11).

Let's now dwell on the second part of the Lemma. For L large enough, as $f_0 > \rho > 0$, then $h > \rho/2 > 0$ by continuity of h on $[-m2^{-L}; 1+m2^{-L}]$. Therefore, Lemma 9 also ensures the existence of a constant $c(\rho)$ such that $\theta_k > c(\rho) > 0,\ m+1 \leq k \leq 2^L+2m$ in (2.25) for L large enough. From (2.23), Lemma 5 and for $\omega_{m,l}$ as in (2.5), we see that

$$2^L \int_0^1 \sum_{k=m+1}^{2^L+2m} \theta_k B_{k,m+1}(t)dt = \sum_{i=2m+1}^{2^L+m} \theta_i + \sum_{i=m+1}^{2m} \left(\omega_{m,i-m}\theta_i + \omega_{m,2m+1-i}\theta_{i+2^L}\right) =: \Omega_m.$$

For f_0 a density, by integration on $[0;1)$, the inequality (2.25) gives

$$\left|2^{-L}\Omega_m - 1\right| \leq C2^{-\alpha L}\left(\left\|f_0^{(\lfloor\alpha\rfloor)}\right\|_\infty + \|f_0\|_{\Sigma(\alpha)}\right).$$

Define

$$\widetilde{\theta}_i := \frac{\theta_i}{2^{-L}\Omega_m}.$$

The two last displays ensure that the $\widetilde{\theta}_i$'s are all bounded by $2\left(\|f_0\|_{\Sigma(\alpha)} + 2\left\|f_0^{(\lfloor\alpha\rfloor)}\right\|_\infty\right)$ for L large enough. From this inequality and (2.25), we now write, for a constant C depending on m, α, $\left\|f_0^{(\lfloor\alpha\rfloor)}\right\|_\infty$ and $\|f_0\|_{\Sigma(\alpha)}$, that

$$\begin{aligned}\left\|f_0 - \sum_{k=m+1}^{2^L+2m}\widetilde{\theta}_k B_{k,m+1}\Big|_{[0;1)}\right\|_\infty &\leq \left\|f_0 - \sum_{k=m+1}^{2^L+2m}\theta_k B_{k,m+1}\Big|_{[0;1)}\right\|_\infty + \\ &\quad \left|1-\left(2^{-L}\Omega_m\right)^{-1}\right|\left\|\sum_{k=m+1}^{2^L+2m}\theta_k B_{k,m+1}\Big|_{[0;1)}\right\|_\infty \\ &\leq C2^{-\alpha L} \qquad (2.26)\end{aligned}$$

since f_0 is bounded on $[0;1)$ as a Hölderian density, the bound depending on α and $\|f\|_{\Sigma(\alpha)}$ only (see [157], p. 9). Let's now define

$$\Theta_k := \begin{cases}\widetilde{\theta}_k & \text{if } 2m+1 \leq k \leq 2^L+m,\\ \omega_{m,k-m}\widetilde{\theta}_k + (1-\omega_{m,k-m})\widetilde{\theta}_{2^L+k} & \text{if } m+1 \leq k \leq 2m,\\ \omega_{m,2^L+2m+1-k}\widetilde{\theta}_k + (1-\omega_{m,2^L+2m+1-k})\widetilde{\theta}_{k-2^L} & \text{if } 2^L+m+1 \leq k \leq 2^L+2m.\end{cases}$$

As pointed out in Lemma 5, $\omega_{m,l} = 1-\omega_{m,m+1-l}$ and, as a consequence, for $m+1 \leq k \leq 2m$, $\Theta_k = \Theta_{2^L+k}$. Then, by definition,

$$\left(2^{-L}\Theta_k\right)_{2m+1\leq k\leq 2^L+2m} \in S^{2^L}.$$

It remains to introduce the application $\sum_{i=0}^{2^L-1} 2^{-L}\Theta_{i+2m+1}S_{i,2^L,m}$ which satisfies, using once again the small support of B-splines (see also Section 2.6.1), that

$$\sum_{i=0}^{2^L-1} 2^{-L}\Theta_{i+2m+1}S_{i,2^L,m}\Big|_{[2^{-L}m;1-2^{-L}m)} = \sum_{k=m+1}^{2^L+2m}\widetilde{\theta}_k B_{k,m+1}\Big|_{[2^{-L}m;1-2^{-L}m)},$$

which, along equation (2.26), finally brings about the conclusion

$$\left\|f_0\big|_{[2^{-L}m;1-2^{-L}m)} - \sum_{i=0}^{2^L-1} 2^{-L}\Theta_{i+2m+1}S_{i,2^L,m}\Big|_{[2^{-L}m;1-2^{-L}m)}\right\|_\infty \leq C2^{-\alpha L}.$$

Therefore, $g = \sum_{i=0}^{2^L-1} 2^{-L}\Theta_{i+2m+1}S_{i,2^L,m}$ satisfies all the conditions from the lemma.

□

2.6 Supplementary results.

We present here additional elements used in the derivation of the main results in the above sections. First, we present results on iterated convolutions of $\mathbb{1}_{[0;1)}$ and spline functions. At the end of this section, we expound on the link between these functions and recall a classic result in Approximation Theory for splines. Indeed, our priors involve iterated convolutions in their definitions, so that a spline approximation theory is helpful here. This allows us to obtain simpler conditions for the derivation of posterior contraction rates for the prior presented in the article. This is the object of additional lemmas that build on a classical result from [64] that we recall, followed by the proof of Theorem 4. Also available, a numerical study of our results presents simulations, focusing on a simplified version of the priors we introduced. We end with some technical results and the presentation of a more flexible extension of DPA.

2.6.1 Results on iterated convolutions of the indicator function and spline functions.

Iterated convolution of the indicator function.

Lemma 2. *For $m \in \mathbb{N}^*$, $\|\chi^{*m}\|_\infty \leq 1$ and*

$$\forall x \notin [0; m], \quad \chi^{*m}(x) = 0.$$

Proof. For $m = 1$, it is straightforward. Then, by induction, from the positivity of χ, for any $x \in \mathbb{R}$,

$$\begin{aligned} 0 \leq \chi^{*(m+1)}(x) &= \int_{\mathbb{R}} \chi^{*m}(t)\chi(x-t)dt \\ &\leq \int_{\mathbb{R}} \mathbb{1}_{[0;m]}(t)\mathbb{1}_{[x-1;x]}(t)dt \\ &\leq \mathbb{1}_{]0;m+1[}(x) \end{aligned}$$

which concludes the proof. □

Lemma 3. *Let $m \in \mathbb{N}^*$ and $s > 0$. Then*

$$\chi_s^{*m} = s^{m-1}\chi^{*m}(\cdot/s).$$

Proof. The result is straightforward for $m = 1$. Now, by induction, assuming that the property for a given m is proved, for $t \in \mathbb{R}$,

$$\begin{aligned} \chi_s^{*(m+1)}(t) &= \int_{\mathbb{R}} \chi_s(u)\, \chi_s^{*m}(t-u)\, du \\ &= s^{m-1} \int_{\mathbb{R}} \chi_s(u)\, \chi^{*m}\left(\frac{t-u}{s}\right) du \\ &= s^m \int_{\mathbb{R}} \chi(v)\, \chi^{*m}\left(\frac{t}{s} - v\right) dv \quad \text{with } v = u/s \\ &= s^m \chi^{*(m+1)}\left(\frac{t}{s}\right). \end{aligned}$$

□

Lemma 4. *Let $m \in \mathbb{N}^*$ and $s > 0$. Then*

$$\int_{\mathbb{R}} \chi^{*m}(t)dt = 1 \qquad \text{and} \qquad s^{-m} \int_{\mathbb{R}} \chi_s^{*m}(t)dt = 1.$$

Proof. The result is straightforward for $m = 1$. For $m \geq 2$, by positivity of χ and with the change of variable $v = t - u$,

$$\int_{\mathbb{R}} \chi^{*m}(t)dt = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} \chi(u)\chi^{*(m-1)}(t-u)du \right] dt = \int_{\mathbb{R}} \chi(u) \left[\int_{\mathbb{R}} \chi^{*(m-1)}(v)dv \right] du = 1$$

by induction. Then, from Lemma 3 and with the change of variable $u = t/s$

$$s^{-m} \int_{\mathbb{R}} \chi_s^{*m}(t)dt = s^{-1} \int_{\mathbb{R}} \chi^{*m}(t/s)dt = \int_{\mathbb{R}} \chi^{*m}(u)du = 1.$$

□

Lemma 5. *Let $m \in \mathbb{N}^*$ and $t \in \mathbb{R}$. Then $\chi^{*m}(m - t) = \chi^{*m}(t)$. As a consequence, for $0 \leq l \leq m + 1$, $\omega_{m,l} = 1 - \omega_{m,m+1-l}$.*

Proof. For $m = 1$, $\chi(1 - t) = \mathbb{1}_{0 \leq 1-t \leq 1} = \mathbb{1}_{0 \leq t \leq 1} = \chi(t)$. By induction, if the theorem is valid until $m \in \mathbb{N}^*$,

$$\begin{aligned}
\chi^{*(m+1)}(m + 1 - t) &= \int_R \chi(u)\chi^{*m}(m + 1 - t - u)du \\
&= \int_{\mathbb{R}} \chi(1 - v)\chi^{*m}(v + m - t)dv \quad \text{with the change of variable } v = 1 - u \\
&= \int_{\mathbb{R}} \chi(v)\chi^{*m}(t - v)dv \\
&= \chi^{*(m+1)}(t).
\end{aligned}$$

The second part of the Lemma comes from a change of variable in (2.5) and Lemma 4.

□

Splines periodicity.

Lemma 6. *Let $(\theta_i)_{i\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}$, $h = 1/N$, $N \in \mathbb{N}^*$ and $m \in \mathbb{N}^*$. The map*

$$x \to \sum_{i\in\mathbb{Z}} \theta_i \chi^{*(m+1)} \left(\frac{x}{h} - i \right)$$

is 1-periodic if and only if the sequence $(\theta_i)_{i\in\mathbb{Z}}$ is N-periodic.

Proof. With $r \in \mathbb{R}$ and $p \in \mathbb{Z}$,

$$\sum_{i\in\mathbb{Z}} \theta_i \chi^{*(m+1)} \left(\frac{r+p}{h} - i \right) = \sum_{j\in\mathbb{Z}} \theta_{j+pN} \chi^{*(m+1)} \left(\frac{r}{h} - j \right).$$

As $\left\{\chi^{(m+1)}\left(h^{-1}\cdot - i\right)\middle| i\in\mathbb{Z}\right\}$ is a basis for the space $\Pi_{m+1,h\mathbb{Z}}(\mathbb{R})$ (see Section 2.5.1), the above map satisfies

$$\sum_{i\in\mathbb{Z}}\theta_i\chi^{*(m+1)}\left(\frac{\cdot}{h}-i\right) \text{ is 1-periodic } \iff \theta_i=\theta_{i+pN} \text{ for any } i\in\mathbb{Z} \text{ and } p\in\mathbb{Z}.$$

□

Lemma 7. *Let $m\in\mathbb{N}$, $L\in\mathbb{N}$ and $(\Theta_i)_{i=0}^{2^L-1}$ be positive real numbers such that*

$$\sum_{i=0}^{2^L-1}\Theta_i=1$$

and introduce, for $i=2^L-m,\dots,2^L-1$,

$$\theta_i\in\left[0;\frac{\Theta_i}{\omega_{m,2^L-i}}\right]$$

where $\omega_{m,i}$ is defined by (2.5). Then, if $(u_i)_{i\in\mathbb{Z}}$ is an (2^L+m)-periodic sequence, such that

$$u_i=\begin{cases}\Theta_i & \text{if } 0\le i\le 2^L-m-1\\ \theta_i & \text{if } 2^L-m\le i\le 2^L-1\\ \frac{\Theta_{i-m}-\omega_{m,2^L+m-i}\theta_{i-m}}{1-\omega_{m,2^L+m-i}} & \text{if } 2^L\le i\le 2^L+m-1\end{cases},$$

the restriction of the map $f^m_{\infty,2^{-L}}$ on $[0;1]$, with $f=\sum_{i\in\mathbb{Z}}u_iH_{Li}$, is a probability density function on $[0;1]$.

Proof. From their definition, the u_i's are positive real numbers such that f and $f^m_{\infty,2^{-L}}$ are themselves positive as it can be seen in (2.4). It remains to compute the integral. Beforehand, we recall that for any $n\ge 1$, χ^{*n} is supported on $[0;n]$, so that $\chi^{*n}(2^L\cdot -i)$ is supported on $[2^{-L}i;2^{-L}(i+n)]$, for any $i\in\mathbb{Z}$. The intersection of this last interval with $[0;1]$ has its interior non-empty if and only if $-n+1\le i\le 2^L-1$. Hence,

$$\begin{aligned}
\int_{[0;1]}f^m_{\infty,2^{-L}}(v)dv &= \int_{[0;1]}\sum_{i\in\mathbb{Z}}u_i2^L\chi^{*(m+1)}\left(2^Lv-i\right)dv \quad \text{according to (2.9) and (2.10),}\\
&= \sum_{i=-m}^{2^L-1}u_i2^L\int_{[0;1]}\chi^{*(m+1)}\left(2^Lv-i\right)dv \quad \text{according to the above remark,}\\
&= \sum_{i=-m}^{2^L-1}u_i\int_{[-i;2^L-i]}\chi^{*(m+1)}(r)\,dr \quad \text{with } r=2^Lv-i,\\
&= \sum_{i=-m}^{-1}u_i\int_{-i}^{m+1}\chi^{*(m+1)}(r)\,dr+\sum_{i=0}^{2^L-m-1}u_i\int_0^{m+1}\chi^{*(m+1)}(r)\,dr\\
&\quad+\sum_{i=2^L-m}^{2^L-1}u_i\int_0^{2^L-i}\chi^{*(m+1)}(r)\,dr\\
&\qquad \text{according to the above discussion on the supports,}\\
&= \sum_{i=-m}^{-1}u_i(1-\omega_{m,-i})+\sum_{i=0}^{2^L-m-1}u_i+\sum_{i=2^L-m}^{2^L-1}u_i\omega_{m,2^L-i} \quad \text{using Lemma 4,}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=0}^{2^L-m-1} u_i + \sum_{i=2^L-m}^{2^L-1} u_i \omega_{m,2^L-i} + \sum_{i=2^L}^{2^L+m-1} u_i (1-\omega_{m,2^L+m-i}) \quad \text{by periodicity,} \\
&= \sum_{i=0}^{2^L-m-1} \Theta_i + \sum_{i=2^L-m}^{2^L-1} \theta_i \omega_{m,2^L-i} + \\
&\quad \sum_{i=2^L}^{2^L+m-1} (1-\omega_{m,2^L+m-i}) \frac{\Theta_{i-m} - \omega_{m,2^L+m-i}\theta_{i-m}}{1-\omega_{m,2^L+m-i}} \\
&= \sum_{i=0}^{2^L-1} \Theta_i + \sum_{i=2^L-m}^{2^L-1} \theta_i \omega_{m,2^L-i} - \sum_{i=2^L}^{2^L+m-1} \omega_{m,2^L+m-i}\theta_{i-m} \\
&= 1 + \sum_{i=2^L-m}^{2^L-1} \theta_i \omega_{m,2^L-i} - \sum_{j=2^L-m}^{2^L-1} \theta_j \omega_{m,2^L-j} \\
&= 1.
\end{aligned}
$$

□

In the above proof, we have also proven that the u_i's from the definition of the DPA and CPA distributions are such that

$$
\sum_{i=0}^{2^L-m-1} u_i + \sum_{i=2^L-m}^{2^L-1} \left(\omega_{m,2^L-i} u_i + \omega_{m,i-(2^L-m-1)} u_{i+m} \right) = 1, \tag{2.27}
$$

where we used Lemmas 4 and 5 to express the last terms of the sum.

Definition of spline functions with iterated convolutions.

Lemma 8. *Let $h > 0$ and $m \in \mathbb{N}$. Then, if $\phi \colon \mathbb{R} \longrightarrow \mathbb{R}$ is a right continuous piecewise constant map with breaks at points jh, $j \in \mathbb{Z}$, then $\phi_{\infty,h}^m$ is a spline function of order $m+1$ with knots $\mathbf{t} = (jh)_{j\in\mathbb{Z}}$.*

Proof. Propositions 6.7.1 and 6.7.2 from [5] show that $\chi^{*(m+1)}$ is a spline of order $m+1$ with knots $i = 0, \dots, m+1$. Also, we write $\phi = \sum_{j\in\mathbb{Z}} \theta_j \mathbb{1}_{[jh;(j+1)h)}(\cdot)$ for some sequence $(\theta_j)_{j\in\mathbb{Z}}$. We conclude the proof with (2.4), Subsection 2.5.1 and

$$
\begin{aligned}
\phi_{\infty,h}^m(x) &= \sum_{j\in\mathbb{Z}} h^{-1}\theta_j \left[\chi^{*m}(h^{-1}\cdot) * \mathbb{1}_{[jh;(j+1)h)} \right](x) \\
&= \sum_{j\in\mathbb{Z}} \theta_j h^{-1} \left[\chi^{*m}(h^{-1}\cdot) * \chi(h^{-1}\cdot - j) \right](x) \\
&= \sum_{j\in\mathbb{Z}} \theta_j h^{-1} \int_{\mathbb{R}} \chi^{*m}(h^{-1}s) \chi\left(h^{-1}(x-s) - j \right) ds \\
&= \sum_{j\in\mathbb{Z}} \theta_j h^{-1} \chi^{*(m+1)}\left(h^{-1}x - j \right) \quad \text{with the change of variable } v = h^{-1}s.
\end{aligned}
$$

□

Spline approximation.

Lemma 9. *Suppose $k \geq \alpha > 0$ and $\mathbf{t}$ is a finite knot sequence of step at most T^{-1}, included in a closed bounded interval $I \subset \mathbb{R}$. There exists a constant C depending only on k and α such that for every $f_0 \in \Sigma(\alpha, I)$ and T large enough, there exists $\theta \in \mathbb{R}^{k+\#\mathbf{t}-1}$ with $\|\theta\|_\infty < \left\|f_0^{(\lfloor\alpha\rfloor)}\right\|_\infty + \|f_0\|_{\Sigma(\alpha)}$ and*

$$\left\| \sum_{i=1}^{k+\#\mathbf{t}-1} \theta_i B_{i,k} - f_0 \right\|_\infty \leq CT^{-\alpha} \left(\left\|f_0^{(\lfloor\alpha\rfloor)}\right\|_\infty + \|f\|_{\Sigma(\alpha)} \right)$$

where the $B_{i,k}$'s form the B-spline basis of $\Pi_{k,\mathbf{t}}(I)$. Furthermore, if f_0 is strictly positive, for T large enough, the vector θ can be chosen to have strictly positive coordinates. The θ_i's can indeed be lower bounded by a strictly positive constant depending on the lower bound on f_0.

Proof. This is Lemma E.4 from [65]. □

Plots.

On Figure 2.5, we see (in the particular case $L = 3$ and $m = 3$) that on the interval $[2^{-L}m; 1 - 2^{-L}m)$, the basis functions $S_{i,2^L,m}$ are equal to the basis functions $B_{i+2m+1,m+1}$ from Figure 2.4.

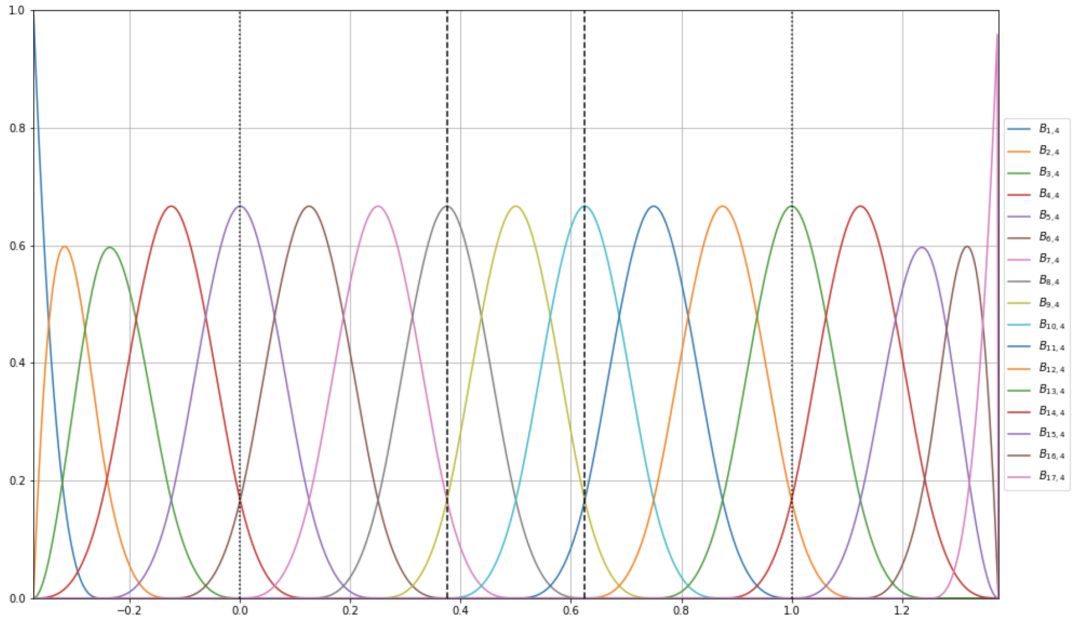


Figure 2.4: B-splines of order 4 with knots $t_i = i/8$ for $-4 \leq i \leq 8 + 4$ as introduced in the proof of Lemma 1.

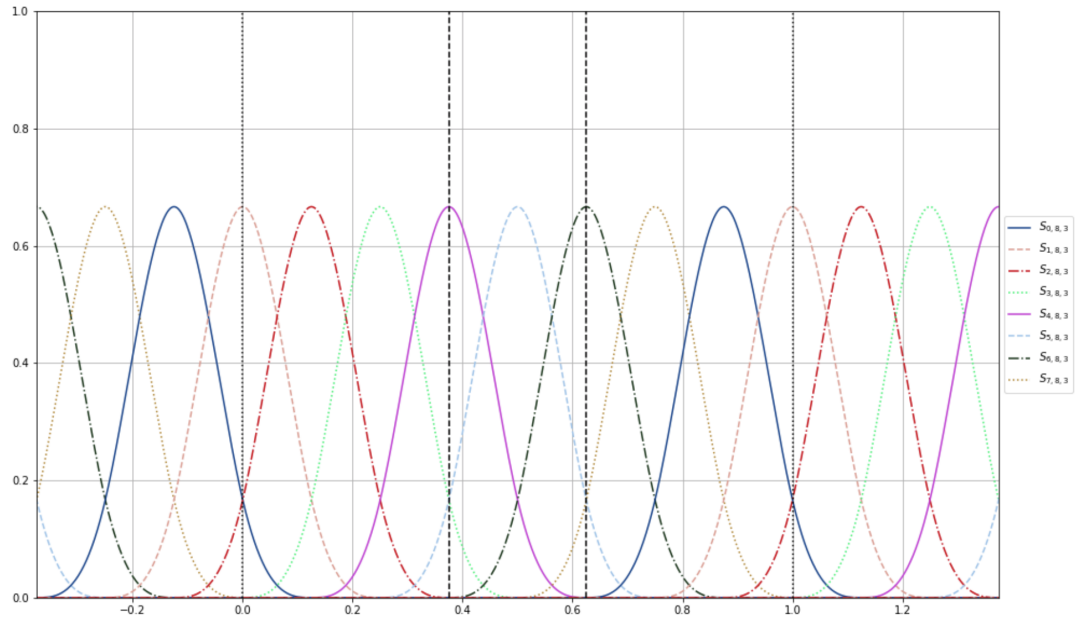


Figure 2.5: Base functions $S_{i,8,3}$.

Figure 2.6 depicts a 1-step aggregation of a TPT sample as introduced in Section 2.2.3. One sees that it smoothes the histogram, as it tends to a piecewise linear density. Even though it remains a histogram function, this added smoothness accounts for better estimation performance with the priors introduced in the before.

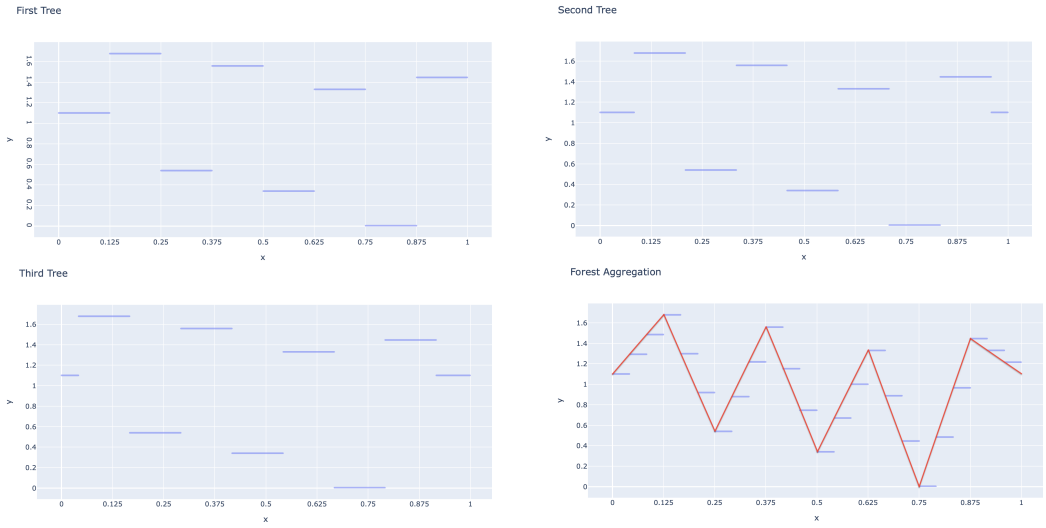


Figure 2.6: "Naïve" aggregation $f^1_{3,2^{-3}}$ where f is the periodic extension of a sample from $\mathrm{TPT}_3(\mathcal{A})$. The first plots represents the shifted trees. The map in red is $f^1_{\infty,2^{-3}}$ when $q \to \infty$.

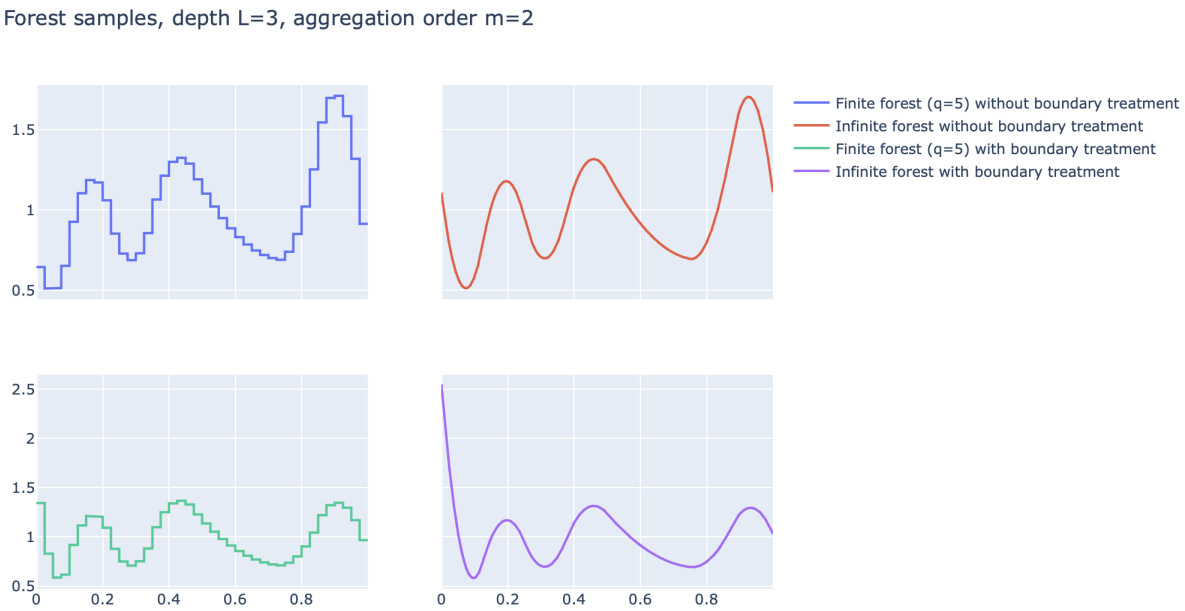


Figure 2.7: Draws from the DPA and CPA priors and their equivalents without the draw of uniform random variable to modify the behaviour near the frontier of $[0; 1)$, with $L = 3$ and $m = 2$.

Figure 2.8: Draws from the DPA and CPA priors and their equivalents without the draw of uniform random variable to modify the behaviour near the frontier of $[0; 1)$, with $L = 6$ and $m = 2$.

2.6.2 Numerical simulations.

Though sampling from the CPA or DPA posterior may be possible via usual MCMC methods, the modification of the samples near the frontier brought by the uniform variables makes

it difficult to explicitly express the posterior or to come up with more efficient sampling algorithms.

However, if we discard the uniform variables from the definition of the prior, it becomes possible to derive an explicit formula of the prior. Namely, for $L > 0$, $q > 0$ and $m \geq 0$, let's focus on the image prior of the $TPT_L(\mathcal{A})$ distribution by the map $f \to f^m_{q,2^{-L}}$ (using definitions from Section 2.2.3). Then, observing an i.i.d. sample $X \in [0;1)^n$, $n > 0$, it is possible to show that the posterior is the image measure by $f \to f^m_{q,2^{-L}}$ of a mixture of TPT distribution, which makes it possible to sample directly from the posterior.

Indeed, for $\mathcal{A} = (a_l)_{0<l\leq L}$ and $Y = \{Y_{\kappa 0},\ 0 \leq |\kappa| < L\}$, we have that the posterior on Y is

$$\Pi[Y|X] \propto f(X_1, \dots, X_n|Y) \prod_{|\kappa|=0}^{L-1} [Y_{\kappa 0}(1 - Y_{\kappa 0})]^{a_{|\kappa|+1}-1},$$

where the likelihood is

$$\begin{aligned} f(X_1, \dots, X_n|Y) &= \prod_{i=1}^{n} \left[q^{-m} \sum_{(j_1,\dots,j_m)\in[\![0;q-1]\!]^m} \sum_{\kappa,|\kappa|=L} \mathbb{1}_{X_i - 2^{-L}(j_1+\dots+j_m)/q \in I_\kappa} 2^L \prod_{j=1}^{L} Y_{\kappa[j]} \right] \\ &= q^{-mn} \sum_{\substack{(j_{1,1},\dots,j_{1,m})\in[\![0;q-1]\!]^m,\dots,\\ (j_{n,1},\dots,j_{n,m})\in[\![0;q-1]\!]^m}} \prod_{i=1}^{n} \sum_{\kappa,|\kappa|=L} \mathbb{1}_{X_i - 2^{-L}(j_1+\dots+j_m)/q \in I_\kappa} 2^L \prod_{j=1}^{L} Y_{\kappa[j]}. \end{aligned}$$

For given $(j_{1,1}, \dots, j_{1,m}), \dots, (j_{n,1}, \dots, j_{n,m})$, all in $[\![0; q-1]\!]^m$, let's note $N_{X,(j_{11},\dots,j_{n,m})}(I_\kappa) = \sum_{i=1}^{n} \mathbb{1}_{X_i - 2^{-L}(j_1+\dots+j_m)/q \in I_\kappa}$ for any $\kappa, |\kappa| \geq 0$, so that

$$\begin{aligned} &\prod_{i=1}^{n} \sum_{\kappa,|\kappa|=L} \mathbb{1}_{X_i - 2^{-L}(j_1+\dots+j_m)/q \in I_\kappa} \prod_{j=1}^{L} Y_{\kappa[j]} = \\ &\qquad \prod_{\kappa, 0\leq|\kappa|\leq L-1} Y_{\kappa 0}^{N_{X,(j_{11},\dots,j_{n,m})}(I_{\kappa 0})} (1 - Y_{\kappa 0})^{N_{X,(j_{11},\dots,j_{n,m})}(I_{\kappa 1})}. \end{aligned}$$

Finally, the posterior on Y is proportional to

$$q^{-mn} \sum_{\substack{(j_{1,1},\dots,j_{1,m})\in[\![0;q-1]\!]^m,\dots,\ \kappa, 0\leq|\kappa|\leq L-1\\ (j_{n,1},\dots,j_{n,m})\in[\![0;q-1]\!]^m}} \prod Y_{\kappa 0}^{N_{X,(j_{11},\dots,j_{n,m})}(I_{\kappa 0})+a_{|\kappa|+1}-1} (1-Y_{\kappa 0})^{N_{X,(j_{11},\dots,j_{n,m})}(I_{\kappa 1})+a_{|\kappa|+1}-1}.$$

The distribution on Y is then a mixture of TPT_L distributions with parameter sets $\mathcal{A}_{(j_{11},\dots,j_{n,m})} = \left\{ N_{X,(j_{11},\dots,j_{n,m})}(I_\kappa) + a_{|\kappa|} - 1,\ 1 \leq |\kappa| \leq L \right\}$ (see [65], Chapter 3).

At the end of this section, we present a theoretical result for this posterior, but we present some simulations first. The density $f_0 : x \in [0;1) \mapsto 1 + 0.5 * \sin(2\pi x)$ satisfies the assumptions of the mentioned theorem with $\alpha = 1.5$. We simulated 10^4 samples from this density and drew 10 samples from the posterior, with parameters tuned as in the below theorem. We also compare these with samples of the TPT posterior with parameters tuned as in [29]. In this paper, sup-norm posterior contraction rates are proven for $\alpha \leq 1$. On Figure 2.9, we can see that the modified DPA prior is associated with smoother posterior samples and leverage the larger regularity of the signal.

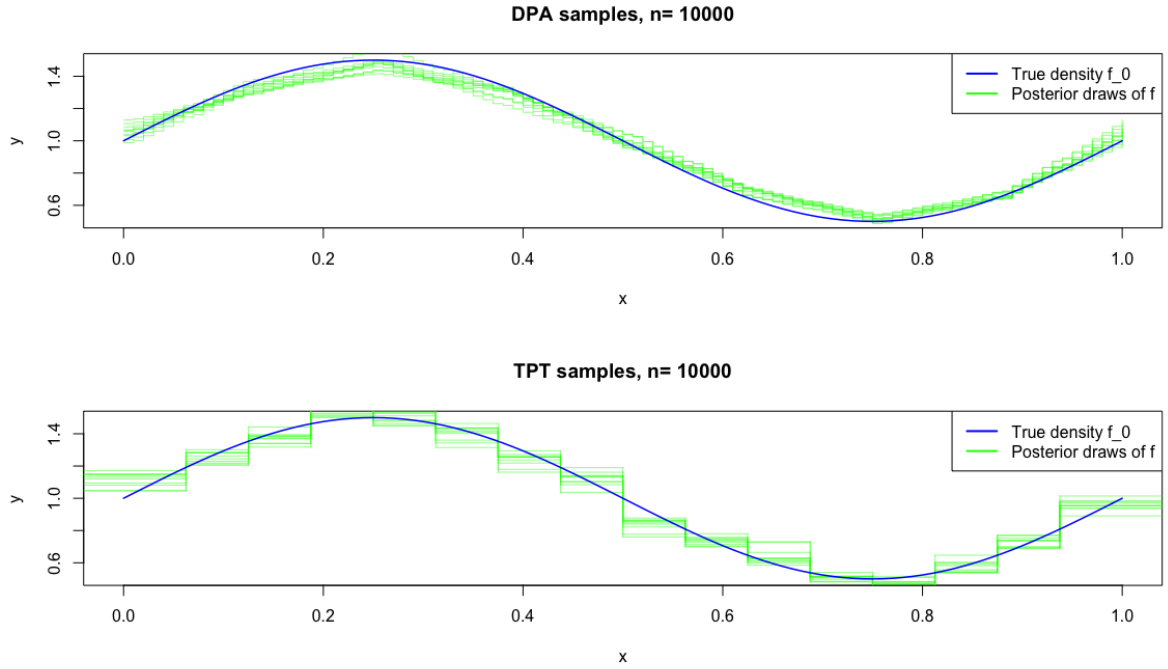


Figure 2.9: Posterior samples for the simplified DPA prior and the TPT prior, with sine sampling density and sample size $n = 10^4$.

We also studied what happens with a density f_0 whose behavior near the frontier renders this simplified prior inadequate. Namely, on Figure 2.10, we analyze the situation where the sampling density f_0 is increasing, $3/2$-smooth, but has different limits towards 0 and 1 (it was obtained as the re-scaled exponential of an integrated Brownian motion). With a sample of size $n = 10^5$ from f_0 and comparing samples from their posteriors, the simplified forest prior still outperforms the single-tree prior far for the frontier. However, as we did not include a modification near the frontier, it behaves badly towards 0 and 1.

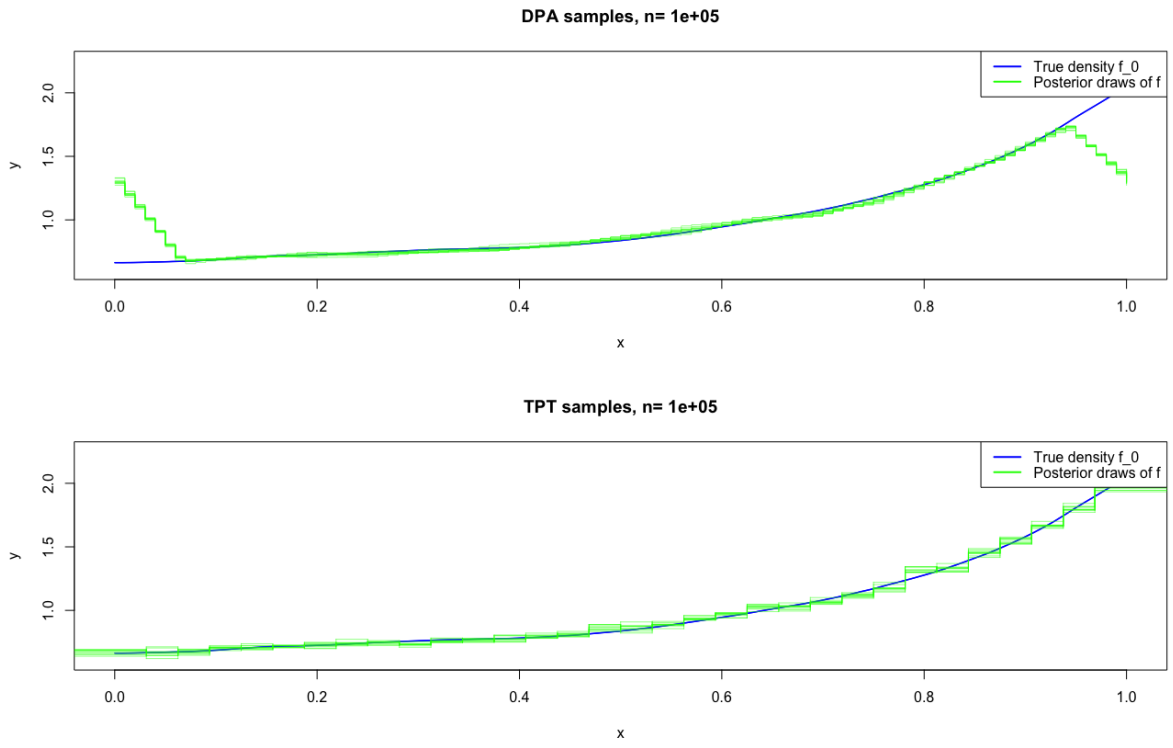


Figure 2.10: Posterior samples for the simplified DPA prior and the TPT prior, with integrated Brownian sampling density and sample size $n = 10^5$.

As discussed in the paper, this prior is not well-suited for the estimation of general smooth densities as it behaves badly near the frontier of Ω. However, if we make additional assumptions on the true densities to be recovered, it is possible to obtain an analog of Theorem 1.

Theorem 6. *Suppose* $f_0 \in \Sigma(\alpha, [0,1))$, $\alpha > 0$, $f_0 \geq \rho$ *for some* $\rho > 0$ *and* $f_0^{(i)}(0) = f_0^{(i)}(1^-)$ *for* $i = 0, \dots, \lfloor \alpha \rfloor$. *If* Π *is the probability distribution of* $f = g^{\lfloor \alpha \rfloor}_{q_n, 2^{-L_n}}$, $g \sim TPT_L\left((a_l)_{0<l\leq L_n}\right)$, *such that for some* $\beta > 0, R \geq 1, \delta > 0$,

- $2^{L_n} \asymp \left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}}$,
- $q_n \geq 2^{\alpha L_n}$,
- $\forall a \in \mathcal{A},\ a \in [r; R]$ *with* $R > r > 0$,

Then, for $M > 0$ *depending on* $\rho, \alpha,\ \|f_0\|_{\Sigma(\alpha)},\ \beta$ *and* R, *and* $d(f,g) = \|f - g\|_1$ *or* $d(f,g) = h(f,g)$, *as* $n \to \infty$,

$$\mathbb{E}_{f_0}\Pi\left[d(f_0, f) > M\left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}} \middle| X\right] \to 0.$$

The proof is similar to the one we provide for DPA in the previous sections, although quite simpler as we do not have to take care of what happens near the frontier.

2.6.3 Contraction rate derivation.

Theorem 7 (Ghosal, Ghosh, Van Der Vaart, 2000). *Suppose d is either the Hellinger or the L_1 distance and Π is an a priori probability distribution on the space of probability densities. Also,*

$$B_{KL}(f_0,\epsilon)=\Bigg\{f:[0,1)\mapsto\mathbb{R}\mid K(f_0,f)\coloneqq\int f_0\log\frac{f_0}{f}\leq\epsilon^2,\\ V(f_0,f)\coloneqq\int f_0\left(\log\frac{f_0}{f}-K(f_0,f)\right)^2\leq\epsilon^2\Bigg\}.$$

If the positive sequence $(\epsilon_n)_{n\geq 0}$ satisfies $\epsilon_n\to 0$, $n\epsilon_n^2\to+\infty$ and there exist sets $\mathcal{F}_n$ such that the three following conditions are satisfied for some $c>0$, $D>0$

1. $\Pi[B_{KL}(f_0,\epsilon_n)]\geq e^{-cn\epsilon_n^2}$,
2. $\log N(\epsilon_n,\mathcal{F}_n,d)\leq Dn\epsilon_n^2$,
3. $\Pi[\mathcal{F}_n^c]\leq e^{-(c+4)n\epsilon_n^2}$,

it then follows, for a constant $M>0$ sufficiently large, depending on c and D, that the posterior satisfies, as $n\to\infty$,

$$\mathbb{E}_{f_0}\Pi[d(f_0,f)>M\epsilon_n|X]\longrightarrow 0.$$

2.6.4 Forest priors DPA and CPA.

Bounds on the Kullback-Leibler divergence.

Lemma 10. *Suppose $f_0\in\Sigma(\alpha,[0,1))$, $\alpha>0$, and $f_0\geq\rho$ for some $\rho>0$. For $m\geq\lfloor\alpha\rfloor$, take $(\eta_i)_{0\leq i\leq 2^L+m+1}$ as the sequence from Lemma 1. Then, for $(u_i)_{i\in\mathbb{Z}}$ a 2^L+m-periodic sequence satisfying (2.27), $L\in\mathbb{N}$,*

$$f=\sum_{i\in\mathbb{Z}}u_iH_{Li},$$

we have that there exists a constant C depending only on ρ, m, α and $\|f_0\|_{\Sigma(\alpha)}$ such that, for $m\geq\lfloor\alpha\rfloor$,

$$K\left(f_0,f^m_{\infty,2^{-L}}\Big|_{[0;1)}\right)\vee V\left(f_0,f^m_{\infty,2^{-L}}\Big|_{[0;1)}\right)\leq C\left(2^{-2\alpha L}+2^{2L}\max_{0\leq i\leq 2^L+m-1}|\eta_i-u_{i-m}|^2\right)$$

for L large enough and $2^L\max_{0\leq i\leq 2^L+m-1}|\eta_i-u_{i-m}|$ small enough. Also, if q is a large enough integer,

$$V\left(f_0,\frac{f^m_{q,2^{-L}}}{\int_0^1 f^m_{q,2^{-L}}(t)dt}\Big|_{[0;1)}\right)\leq C\left(2^{-2\alpha L}+2^{2L}\max_{0\leq i\leq 2^L+m-1}|\eta_i-u_{i-m}|^2+\left(\frac{m\omega_{m,1}^{-1}2^L}{q}\right)^2\right)$$

and

$$K\left(f_0,\frac{f^m_{q,2^{-L}}}{\int_0^1 f^m_{q,2^{-L}}(t)dt}\Big|_{[0;1)}\right)\leq C\left(2^{-2\alpha L}+2^{2L}\max_{0\leq i\leq 2^L+m-1}|\eta_i-u_{i-m}|^2+\left(\frac{m\omega_{m,1}^{-1}2^L}{q}\right)^2\right).$$

Proof. From Lemma 1, the map $\tilde{f}_0^L = \sum_{i=0}^{2^L+m-1} \eta_i 2^L \chi^{*(m+1)}\left(2^L \cdot -(i-m)\right)$ is such that

$$\left\| f_0 - \tilde{f}_0^L\big|_{[0;1)} \right\|_\infty \leq C 2^{-\alpha L} \tag{2.28}$$

with C depending on m, α and $\|f_0\|_{\Sigma(\alpha)}$. Let's first write the decomposition, given A and B exist (it will be shown later),

$$\begin{aligned} K\left(f_0, f^m_{\infty,2^{-L}}\big|_{[0;1)}\right) &= \int_0^1 f_0(t) \log\left(\frac{f_0(t)}{f^m_{\infty,2^{-L}}(t)}\right) dt \\ &= \underbrace{\int_0^1 f_0(t) \log\left(\frac{f_0(t)}{\tilde{f}_0^L(t)}\right) dt}_{=\mathsf{A}} + \underbrace{\int_0^1 f_0(t) \log\left(\frac{\tilde{f}_0^L(t)}{f^m_{\infty,2^{-L}}(t)}\right) dt}_{=\mathsf{B}}. \end{aligned}$$

Focusing on the first term, we have, as $\log(1+u) \leq u$ for $u > -1$,

$$\begin{aligned} A &\leq \int_0^1 f_0(t) \frac{f_0(t) - \tilde{f}_0^L(t)}{\tilde{f}_0^{L(t)}} dt \\ &= \int_0^1 \frac{\left(f_0(t) - \tilde{f}_0^L(t)\right)^2}{\tilde{f}_0^{L(t)}} dt + 1 - \int_0^1 \tilde{f}_0^L(t) dt \\ &= \int_0^1 \frac{\left(f_0(t) - \tilde{f}_0^L(t)\right)^2}{\tilde{f}_0^{L(t)}} dt + 1 - \left[\sum_{k=m}^{2^L-1} \eta_k + \sum_{k=0}^{m-1} \left(\omega_{m,k+1}\eta_k + (1-\omega_{m,m-k})\eta_{2^L+k}\right)\right] \\ &\leq \frac{2}{\rho} \left\| f_0 - \tilde{f}_0^L\big|_{[0;1)} \right\|_\infty^2 \quad \text{for } L \text{ large enough,} \\ &\leq \frac{C}{\rho} 2^{-2\alpha L}, \end{aligned}$$

where we lower bounded $\tilde{f}_0^L$ by $\rho/2$ on $[0;1)$ for L large enough as f_0 is lower bounded by $\rho > 0$. On the other hand, since $\tilde{f}_0^L$ and $f^m_{\infty,2^{-L}}$ have unit integral on $[0;1)$ according to their definitions, we have the upper bound

$$\begin{aligned} B &\leq \int_0^1 f_0(t) \frac{\tilde{f}_0^L(t) - f^m_{\infty,2^{-L}}(t)}{f^m_{\infty,2^{-L}}(t)} dt \\ &= \int_0^1 \frac{\left(\tilde{f}_0^L(t) - f^m_{\infty,2^{-L}}(t)\right)\left(f_0(t) - f^m_{\infty,2^{-L}}(t)\right)}{f^m_{\infty,2^{-L}}(t)} dt \\ &= \int_0^1 \frac{\left(\tilde{f}_0^L(t) - f^m_{\infty,2^{-L}}(t)\right)\left(f_0(t) - \tilde{f}_0^L(t)\right)}{f^m_{\infty,2^{-L}}(t)} dt + \int_0^1 \frac{\left(\tilde{f}_0^L(t) - f^m_{\infty,2^{-L}}(t)\right)^2}{f^m_{\infty,2^{-L}}(t)} dt \\ &\leq \frac{1}{2}\int_0^1 \frac{\left(\tilde{f}_0^L(t) - f_0(t)(t)\right)^2}{f^m_{\infty,2^{-L}}(t)} dt + \frac{3}{2}\int_0^1 \frac{\left(\tilde{f}_0^L(t) - f^m_{\infty,2^{-L}}(t)\right)^2}{f^m_{\infty,2^{-L}}(t)} dt \text{ using that } 2ab \leq a^2 + b^2 \text{ for real numbers.} \end{aligned}$$

At this point, we develop from (2.10) and the fact that the maps $\chi^{*(m+1)}(2^L \cdot -k)$ make a partition of the unity (see section E.2 from [65])

$$\begin{aligned} \left\| \tilde{f}_0^L\big|_{[0;1)} - f^m_{\infty,2^{-L}}\big|_{[0;1)} \right\|_\infty &\leq \left\| \sum_{i=0}^{2^L+m-1} (\eta_i - u_{i-m}) 2^L \chi^{*(m+1)}(2^L \cdot -(i-m)) \right\|_\infty \\ &\leq 2^L \max_{0 \leq i \leq 2^L+m-1} |\eta_i - u_{i-m}|. \end{aligned} \tag{2.29}$$

Therefore, for L large enough and $2^L \max_{0 \leq i \leq 2^L+m-1} |\eta_i - u_{i-m}|$ small enough, we also lower bound $f^m_{\infty,2^{-L}}$ by $\rho/4$. Under such conditions, we finally give the bounds

$$B \leq \frac{2}{\rho} \left\| f_0 - \tilde{f}_0^L\big|_{[0;1)} \right\|_\infty^2 + \frac{6}{\rho} \left\| \tilde{f}_0^L - f^m_{\infty,2^{-L}} \right\|_\infty^2 \leq C\left(2^{-2\alpha L} + 2^{2L} \max_{0 \leq i \leq 2^L+m-1} |\eta_i - u_{i-m}|^2\right)$$

and

$$K(f_0, f^m_{\infty,2^{-L}}) \leq C\left(2^{-2\alpha L} + 2^{2L} \max_{0\leq i\leq 2^L+m-1} |\eta_i - u_{i-m}|^2\right).$$

For the discrete version,

$$K\left(f_0, \frac{f^m_{q,2^{-L}}}{\int_0^1 f^m_{q,2^{-L}}(t)dt}\Big|_{[0;1)}\right) \leq K(f_0, f^m_{\infty,2^{-L}}) + \int_{[0;1)} f_0 \log\left(\frac{f^m_{\infty,2^{-L}}}{f^m_{q,2^{-L}}\left(\int_0^1 f^m_{q,2^{-L}}(t)dt\right)^{-1}}\right) d\lambda.$$

It then remains to use Lemma 17 to see that for q large enough, since $0 \leq f \leq 2^L\omega^{-1}_{m,1}$ from (2.27) and $\inf_{1\leq l\leq m} \omega_{m,l} = \omega_{m,1}$,

$$\begin{aligned}
&\int_{[0;1)} f_0 \log\left(\frac{f^m_{\infty,2^{-L}}}{f^m_{q,2^{-L}}\left(\int_0^1 f^m_{q,2^{-L}}(t)dt\right)^{-1}}\right) d\lambda \\
&\leq \int_{[0;1)} f_0 \frac{\left(f^m_{\infty,2^{-L}} - f^m_{q,2^{-L}}\left(\int_0^1 f^m_{q,2^{-L}}(t)dt\right)^{-1}\right)}{f^m_{q,2^{-L}}\left(\int_0^1 f^m_{q,2^{-L}}(t)dt\right)^{-1}} \\
&= \int_{[0;1)} \left(f_0 - \tilde{f}^L_0 + \tilde{f}^L_0 - f^m_{\infty,2^{-L}} + f^m_{\infty,2^{-L}} - f^m_{q,2^{-L}}\left(\int_0^1 f^m_{q,2^{-L}}(t)dt\right)^{-1}\right) \\
&\qquad\qquad \frac{\left(f^m_{\infty,2^{-L}} - f^m_{q,2^{-L}}\left(\int_0^1 f^m_{q,2^{-L}}(t)dt\right)^{-1}\right)}{f^m_{q,2^{-L}}\left(\int_0^1 f^m_{q,2^{-L}}(t)dt\right)^{-1}} d\lambda \\
&\leq \frac{4}{\rho}\left\|f_0 - \tilde{f}^L_0\right\|^2_\infty + \frac{4}{\rho}\left\|\tilde{f}^L_0 - f^m_{\infty,2^{-L}}\right\|^2_\infty + \frac{16}{\rho}\left\|f^m_{\infty,2^{-L}} - f^m_{q,2^{-L}}\left(\int_0^1 f^m_{q,2^{-L}}(t)dt\right)^{-1}\right\|^2_\infty \\
&\quad \text{since } 2ab \leq a^2+b^2 \text{ and } (a+b+c)^2 \leq 3(a^2+b^2+c^2), \\
&\leq C\left(2^{-2\alpha L} + 2^{2L}\max_{0\leq i\leq 2^L+m-1} |\eta_i - u_{i-m}|^2 + \left(\frac{m\omega^{-1}_{m,1}2^L}{q}\right)^2\right),
\end{aligned}$$

where we used that for L, q large enough and $2^L \max_{0\leq i\leq 2^L+m-1} |\eta_i - u_{i-m}|$ small enough,

$$f^m_{q,2^{-L}}\left(\int_0^1 f^m_{q,2^{-L}}(t)dt\right)^{-1} \geq \rho/8.$$

On the other hand, f_0 belongs to an interval of the form $[\rho/2; M]$ as f_0 is upper bounded by a constant depending on α and $\|f_0\|_{\Sigma(\alpha)}$ only since it is a Hölderian density (see [157], p.9). Also, from equations (2.28) and (2.29), $f^m_{\infty,2^{-L}} \in [\rho/4; 2M]$ for L large enough and $2^L \max_{0\leq i\leq 2^L+m-1} |\eta_i - u_{i-m}|$ small enough. We then use Taylor's inequality to write that

$$\begin{aligned}
V(f_0, f^m_{\infty,2^{-L}}) &= \int f_0\left(\log f_0 - \log f^m_{\infty,2^{-L}}\right)^2 d\lambda \\
&\lesssim \left\|f_0 - f^m_{\infty,2^{-L}}\Big|_{[0;1)}\right\|^2_\infty
\end{aligned}$$

as well as

$$V\left(f_0, \frac{f^m_{q,2^{-L}}}{\int_0^1 f^m_{q,2^{-L}}(t)dt}\Big|_{[0;1)}\right) \lesssim \left\|f_0 - \frac{f^m_{q,2^{-L}}}{\int_0^1 f^m_{q,2^{-L}}(t)dt}\Big|_{[0;1)}\right\|^2_\infty$$

from Lemma 17 and q large enough. We conclude with the triangular inequality, equations (2.28), (2.29) and Lemma 17. □

Bounds on the Hellinger distance.

Lemma 11. *Let $(\theta_i)_{i\in\mathbb{Z}}$ and $(\zeta_i)_{i\in\mathbb{Z}}$ be two $\left(2^L+m\right)$-periodic sequences of real positive numbers in $\mathcal{H}_{m,L}$ from (2.13), where $L\in\mathbb{N}^*, m\in\mathbb{N}$ are such that $m<2^L-1$, and verifying*

$$\|\theta\|_\infty \vee \|\zeta\|_\infty \leq M \in \mathbb{R}_+^*.$$

If $f=\sum_{i\in\mathbb{Z}}\theta_i H_{Li}$ and $g=\sum_{i\in\mathbb{Z}}\zeta_i H_{Li}$, then, for $q\in\mathbb{N}^*$, $q>2^{L+1}mM$,

$$h\left(\left.\frac{f^m_{q,2^{-L}}}{\int_0^1 f^m_{q,2^{-L}}(t)dt}\right|_{[0;1)}, \left.\frac{g^m_{q,2^{-L}}}{\int_0^1 g^m_{q,2^{-L}}(t)dt}\right|_{[0;1)}\right)$$
$$\leq 3^{1/4}\left(2^L+m\right)^{1/4}\|\theta-\zeta\|_2^{1/2} + 6^{1/4}\sqrt{\left(\frac{2^{L+1}Mm}{q-2^{L+1}Mm}\right)(1+mM)},$$

as well as

$$h\left(\left.f^m_{\infty,2^{-L}}\right|_{[0;1)}, \left.g^m_{\infty,2^{-L}}\right|_{[0;1)}\right) \leq \left(2^L+m\right)^{1/4}\|\theta-\zeta\|_2^{1/2}.$$

Proof. First, the same computation as in the proof of Lemma 7 shows that

$$\int_0^1 f^m_{\infty,2^{-L}}(t)dt = 1.$$

Then, according to Lemma 17, there exists V such that $|V|\leq \frac{2^{L+1}Mm}{q}$ and

$$\int_0^1 f^m_{q,2^{-L}}(t)dt = 1+V$$

as f takes values in $[0;2^LM]$. The same properties are verified with g, which allows us to write

$$h\left(\left.\frac{f^m_{q,2^{-L}}}{\int_0^1 f^m_{q,2^{-L}}(t)dt}\right|_{[0;1)}, \left.\frac{g^m_{q,2^{-L}}}{\int_0^1 g^m_{q,2^{-L}}(t)dt}\right|_{[0;1)}\right)$$
$$= \left(\int_0^1\left(\sqrt{\frac{1}{(1+V)q^m}\sum_{(i_1,\dots,i_m)\in[\![0;q-1]\!]^m} f\left(u-\frac{i_1+\dots+i_m}{q2^L}\right)}\right.\right.$$
$$\left.\left. - \sqrt{\frac{1}{(1+V')q^m}\sum_{(i_1,\dots,i_m)\in[\![0;q-1]\!]^m} g\left(u-\frac{i_1+\dots+i_m}{q2^L}\right)}\right)^2 du\right)^{1/2}$$
$$\leq \left(\int_0^1\left|\sum_{(i_1,\dots,i_m)\in[\![0;q-1]\!]^m}\frac{f\left(u-\frac{i_1+\dots+i_m}{q2^L}\right)}{(1+V)q^m} - \frac{g\left(u-\frac{i_1+\dots+i_m}{q2^L}\right)}{(1+V')q^m}\right| du\right)^{1/2}$$
$$\leq \left(\frac{1}{q^m}\sum_{(i_1,\dots,i_m)\in[\![0;q-1]\!]^m}\int_{-\frac{i_1+\dots+i_m}{q2^L}}^{1-\frac{i_1+\dots+i_m}{q2^L}}\left|\frac{f\left(u\right)}{(1+V)} - \frac{g\left(u\right)}{(1+V')}\right| du\right)^{1/2}$$

$$
\begin{aligned}
&\leq \left(\int_{-m2^{-L}}^{1} \left| \frac{f\big(u\big)}{(1+V)} - \frac{g\big(u\big)}{(1+V')} \right| du \right)^{1/2} \\
&\leq \left((1+m2^{-L})^{1/2} \left[\int_{-m2^{-L}}^{1} \left(\frac{f\big(u\big)}{(1+V)} - \frac{g\big(u\big)}{(1+V')} \right)^2 du \right]^{1/2} \right)^{1/2} \\
&= \left((1+m2^{-L})^{1/2} \left[\sum_{i=0}^{2^L+m-1} 2^L \left(\frac{\theta_i}{(1+V)} - \frac{\zeta_i}{(1+V')} \right)^2 \right]^{1/2} \right)^{1/2} \\
&\leq 3^{1/4} 2^{L/4} (1+m2^{-L})^{1/4} \left[\sum_{i=0}^{2^L+m-1} (\theta_i - \zeta_i)^2 + \sum_{i=0}^{2^L+m-1} \left(\frac{V}{1+V} \theta_i \right)^2 + \sum_{i=0}^{2^L+m-1} \left(\frac{V'}{1+V'} \zeta_i \right)^2 \right]^{1/4} \\
&\leq 3^{1/4} 2^{L/4} (1+m2^{-L})^{1/4} \left[\|\theta - \zeta\|_2^2 + \left(\frac{V}{1+V} \right)^2 \left(\sum_{i=0}^{2^L+m-1} \theta_i \right)^2 + \left(\frac{V'}{1+V'} \right)^2 \left(\sum_{i=0}^{2^L+m-1} \zeta_i \right)^2 \right]^{1/4} \\
&\leq 3^{1/4} 2^{L/4} (1+m2^{-L})^{1/4} \left[\|\theta - \zeta\|_2^2 + 2 \left(\frac{2^{L+1} Mm}{q - 2^{L+1} Mm} \right)^2 (1+mM)^2 \right]^{1/4} \\
&= 3^{1/4} \left(2^L + m \right)^{1/4} \|\theta - \zeta\|_2^{1/2} + 6^{1/4} \sqrt{\left(\frac{2^{L+1} Mm}{q - 2^{L+1} Mm} \right) (1+mM)}.
\end{aligned}
$$

Above, we have used that, since $\omega_{m,l} = 1 - \omega_{m,m+1-l}$ from Lemma 4,

$$
\begin{aligned}
\sum_{i=0}^{2^L+m-1} \theta_i &= 1 + \sum_{i=2^L-m}^{2^L-1} \left(\omega_{m,m+1-(2^L-i)} \theta_i + \omega_{m,2^L-i} \theta_{i+m} \right) \\
&\leq 1 + \sum_{i=2^L-m}^{2^L-1} \left(\omega_{m,m+1-(2^L-i)} + \omega_{m,2^L-i} \right) M \\
&\leq 1 + mM.
\end{aligned}
$$

Also,

$$
\begin{aligned}
h\left(f_{\infty,2^{-L}}^m \Big|_{[0;1)}, g_{\infty,2^{-L}}^m \Big|_{[0;1)} \right) &= \left(\int_0^1 \left(\sqrt{\sum_{i\in\mathbb{Z}} \theta_i 2^L \chi^{*m}\left(2^L u - i\right)} - \sqrt{\sum_{i\in\mathbb{Z}} \zeta_i 2^L \chi^{*m}\left(2^L u - i\right)} \right)^2 du \right)^{1/2} \\
&\leq \left(\int_0^1 \left| \sum_{i\in\mathbb{Z}} (\theta_i - \zeta_i) 2^L \chi^{*m}\left(2^L t - i\right) \right| du \right)^{1/2} \\
&\leq \left(\sum_{i\in\mathbb{Z}} |\theta_i - \zeta_i| \, 2^L \int_0^1 \chi^{*m}\left(2^L t - i\right) du \right)^{1/2} \\
&\leq \left(\sum_{i=0}^{2^L+m-1} |\theta_i - \zeta_i| \right)^{1/2} \\
&\leq \left(2^L + m \right)^{1/4} \left(\sum_{i=0}^{q+m-1} |\theta_i - \zeta_i|^2 \right)^{1/4} \\
&\leq \left(2^L + m \right)^{1/4} \|\theta - \zeta\|_2^{1/2}.
\end{aligned}
$$

□

2.6.5 Spline prior SPT.

Bounds on the Kullback-Leibler divergence.

Lemma 12. *Suppose $f_0 \in \Sigma(\alpha, [0,1))$, for $\alpha > 0$, is a probability density and $f_0 \geq \rho$ for some $\rho > 0$. Define $(\eta_i)_{0 \leq i \leq 2^L - 1}$ as the sequence from Lemma 15. Then, for $\tau > 0$ small enough, $L \in \mathbb{N}^*$ large enough, $(\Theta_i)_{0 \leq i \leq 2^L - 1} \in S^{2^L}$ and*

$$f = SD_{\tau, \lfloor \alpha \rfloor, 2^{-L}} \left(\sum_{i=0}^{2^{L_n}-1} \Theta_i H_{Li} \right),$$

we have that there exists a constant C depending only on ρ, α, $\|f_0\|_\infty$ and $\|f_0\|_{\Sigma(\alpha)}$ such that

$$K(f_0, f) \vee V(f_0, f) \leq C \left(2^{-2\alpha L} + \tau^2 + \left(2^L \tau^{-1} \vee 2^{2L} \tau^{-2} \right)^2 \max_{0 \leq i \leq 2^L - 1} |\eta_i - \Theta_i|^2 \right)$$

for $\left(2^L \tau^{-1} \vee 2^{2L} \tau^{-2} \right) \max_{0 \leq i \leq 2^L - 1} |\eta_i - \Theta_i|$ small enough.

Proof. Let f_0^L be the map obtained from Lemma 15, then

$$f_0^L = SD_{\tau, \lfloor \alpha \rfloor, 2^{-L}} \left(\sum_{i=0}^{2^L - 1} \eta_i H_{Li} \right)$$

for some $(\eta_i)_{0 \leq i \leq 2^L - 1} \in S^{2^L}$. We now give the decomposition, given A and B exist (it will be shown later),

$$\begin{aligned} K(f_0, f) &= \int f_0 \log \left(\tfrac{f_0}{f} \right) d\lambda \\ &= \underbrace{\int f_0 \log \left(\frac{f_0}{f_0^L} \right) d\lambda}_{= \mathsf{A}} + \underbrace{\int f_0 \log \left(\frac{f_0^L}{f} \right) d\lambda}_{= \mathsf{B}}. \end{aligned}$$

Focusing on the first term, the bound $\log(1+u) \leq u$ for $u > -1$ results in

$$\begin{aligned} A &\leq \int_0^1 f_0 \tfrac{f_0 - f_0^L}{f_0^L} d\lambda \\ &= \int_0^1 \tfrac{(f_0 - f_0^L)^2}{f_0^L} d\lambda + 1 - \int_0^1 f_0^L d\lambda \\ &\leq \tfrac{2}{\rho} ||f_0 - f_0^L||_\infty^2 \quad \text{for } L \text{ large enough,} \end{aligned} \tag{2.30}$$

where we used that, by construction, $\int_0^1 f_0^L(t) dt = 1$. Also, we have lower bounded f_0^L by $\rho/2$ for L large enough and τ small enough, as a consequence from our assumption on f_0 and Lemma 15. On the other hand, using once again that f_0^L is a density, we have the upper bound

$$\begin{aligned} B &\leq \int_0^1 f_0 \tfrac{f_0^L - f}{f} d\lambda \\ &= \int_0^1 \tfrac{(f_0^L - f)(f_0 - f)}{f} d\lambda \\ &= \int_0^1 \tfrac{(f_0^L - f)(f_0 - f_0^L)}{f} d\lambda + \int_0^1 \tfrac{(f_0^L - f)^2}{f} d\lambda \\ &\leq \tfrac{1}{2} \int_0^1 \tfrac{(f_0 - f_0^L)^2}{f} d\lambda + \tfrac{3}{2} \int_0^1 \tfrac{(f_0^L - f)^2}{f} d\lambda \text{ using that } 2ab \leq a^2 + b^2 \text{ for } a, b \text{ real numbers.} \end{aligned} \tag{2.31}$$

Along with the lower bound on f_0^L, Lemma 14 ensures that there exists a constant C depending on ρ and α only such that, when L is large enough and $\left(2^L\tau^{-1} \vee 2^{2L}\tau^{-2}\right) \max_{0\leq i\leq 2^L-1} |\eta_i - \Theta_i|$ is small enough,

$$B \leq C\left(\left\|f_0 - f_0^L\right\|_\infty^2 + \left(2^L\tau^{-1} \vee 2^{2L}\tau^{-2}\right)^2 \max_{0\leq i\leq 2^L-1} |\eta_i - \Theta_i|^2 \right).$$

In the end, (2.30), (2.31) and Lemma 15 lead to the bound

$$K(f_0, f) \leq C\left(2^{-2\alpha L} + \tau^2 + \left(2^L\tau^{-1} \vee 2^{2L}\tau^{-2}\right)^2 \max_{0\leq i\leq 2^L-1} |\eta_i - \Theta_i|^2\right)$$

where the constant C only depends on α, ρ, $\|f_0\|_\infty$ and $\|f_0\|_{\Sigma(\alpha)}$.

In a second part, we write that

$$\begin{aligned} V(f_0, f) &= \int f_0 \log\left(\tfrac{f_0}{f}\right)^2 d\lambda \\ &\leq \underbrace{2\int f_0 \log\left(\frac{f_0}{f_0^L}\right)^2 d\lambda}_{=\text{A'}} + \underbrace{2\int f_0 \log\left(\frac{f_0^L}{f}\right)^2 d\lambda}_{=\text{B'}}. \end{aligned}$$

For the first term, introducing the Lebesgue-measurable event $G = \left\{x \in [0;1) \;\middle|\; f_0(x) > f_0^L(x)\right\}$, we use similar arguments as above to write that

$$\begin{aligned} A' &= \int_G f_0 \log\left(\tfrac{f_0}{f_0^L}\right)^2 d\lambda + \int_{G^c} f_0 \log\left(\tfrac{f_0^L}{f_0}\right)^2 d\lambda \\ &\leq \int_G f_0 \left(\tfrac{f_0 - f_0^L}{f_0^L}\right)^2 d\lambda + \int_{G^c} f_0 \left(\tfrac{f_0^L - f_0}{f_0}\right)^2 d\lambda \\ &\leq C\left\|f_0 - f_0^L\right\|_\infty^2 \end{aligned} \tag{2.32}$$

where C depends on ρ only and the last inequality is valid for L large enough and τ small enough. Similarly, for the second term, introducing the Lebesgue-measurable event $H = \left\{x \in [0;1) \;\middle|\; f_0^L(x) > f(x)\right\}$ and using Lemma 14, it follows that

$$\begin{aligned} B' &\leq \int_H f_0 \left(\tfrac{f_0 - f}{f}\right)^2 d\lambda + \int_{H^c} f_0 \left(\tfrac{f_0^L - f}{f_0^L}\right)^2 d\lambda \\ &= C\left(2^L\tau^{-1} + 2^{2L}\tau^{-2}\right)^2 \max_{0\leq i\leq 2^L-1} |\eta_i - \theta_i|^2 \end{aligned} \tag{2.33}$$

for some C, which is valid for L large enough and $\left(2^L\tau^{-1} \vee 2^{2L}\tau^{-2}\right) \max_{0\leq i\leq 2^L-1} |\eta_i - \Theta_i|$ small enough. Finally, we obtain from (2.32), (2.33) and Lemma 15

$$V(f_0, f) \leq C\left(2^{-2\alpha L} + \tau^2 + \left(2^L\tau^{-1} \vee 2^{2L}\tau^{-2}\right)^2 \max_{0\leq i\leq 2^L-1} |\eta_i - \Theta_i|^2\right)$$

with C depending on α, ρ, $\|f_0\|_\infty$ and $\|f_0\|_{\Sigma(\alpha)}$. □

Bound on the Hellinger distance.

Lemma 13. *Let $\tau > 0$, $m \in \mathbb{N}$, $L \in \mathbb{N}^*$ and two density functions g_1, g_2 on $[0;1)$ which are piecewise constant on the grid $[i2^{-L}; (i+1)2^{-L})$, $0 \leq i \leq 2^L - 1$. Then, for $f_i = SD_{\tau,m,2^{-L}}(g_i)$,*

$$h(f_1, f_2) \leq 2\left(1 + \sqrt{1 + 2(m+1)^3 e^{\sqrt{6(m+1)}m}}\right) \tau^{-1/2} \|g_1 - g_2\|_2^{1/2}.$$

Proof. By definition,

$$f_i = \frac{A^2_{m,2^{-L}}(g_i)_+ + \tau}{\int_{[0;1)} (A^2_{m,2^{-L}}(g_i)_+ + \tau) d\lambda}.$$

Then the triangle inequality, its reversed version and simple algebra give

$$\begin{aligned} h(f_1, f_2) &= \left\| \frac{\sqrt{A^2_{m,2^{-L}}(g_1)_+ + \tau}}{\left\|\sqrt{A^2_{m,2^{-L}}(g_1)_+ + \tau}\right\|_2} - \frac{\sqrt{A^2_{m,2^{-L}}(g_2)_+ + \tau}}{\left\|\sqrt{A^2_{m,2^{-L}}(g_2)_+ + \tau}\right\|_2} \right\|_2 \\ &\leq 2 \frac{\left\|\sqrt{A^2_{m,2^{-L}}(g_1)_+ + \tau} - \sqrt{A^2_{m,2^{-L}}(g_2)_+ + \tau}\right\|_2}{\left\|\sqrt{A^2_{m,2^{-L}}(g_1)_+ + \tau}\right\|_2}. \end{aligned}$$

The numerator on the right hand side is then bounded by

$$\begin{aligned} \left\|A^2_{m,2^{-L}}(g_1)_+ - A^2_{m,2^{-L}}(g_2)_+\right\|_1^{1/2} &\leq \left\|A^1_{m,2^{-L}}(g_1) - A^1_{m,2^{-L}}(g_2)\right\|_1^{1/2} \\ &+ \left\|\left(A^2_{m,2^{-L}}(g_1)_+ - A^1_{m,2^{-L}}(g_1)\right) - \left(A^2_{m,2^{-L}}(g_2)_+ - A^1_{m,2^{-L}}(g_2)\right)\right\|_1^{1/2} \end{aligned}$$

as for any $a, b \geq 0$, $|\sqrt{a} - \sqrt{b}| \leq \sqrt{|a-b|}$. For the first term, with $\bar{g}_i$ the 1-periodic extension of g_i, we develop

$$\begin{aligned} \left\|A^1_{m,2^{-L}}(g_1) - A^1_{m,2^{-L}}(g_2)\right\|_1 &= \int_0^1 \left|\bar{g}^m_{1,\infty,2^{-L}}(t) - \bar{g}^m_{2,\infty,2^{-L}}(t)\right| dt \\ &= \int_0^1 \left|\chi^m_{2^{-L}} * (\bar{g}_1 - \bar{g}_2)(t)\right| dt \\ &\leq \int_0^1 \chi^m_{2^{-L}} * (|\bar{g}_1 - \bar{g}_2|)(t) dt \\ &= \int_{\mathbb{R}} \chi^m_{2^{-L}}(x) \int_0^1 |\bar{g}_1 - \bar{g}_2|(t-x) dt dx \\ &= \int_{\mathbb{R}} \chi^m_{2^{-L}}(x) \int_0^1 |g_1 - g_2|(t) dt dx \\ &= \|g_1 - g_2\|_1, \end{aligned}$$

following Lemma 4. Then, according to the link between $A^1_{m,2^{-L}}(g_i)$ and $A^2_{m,2^{-L}}(g_i)$, the square of the second term is equal to

$$\begin{aligned} &\int_{\left[0;\frac{m+1}{2^L}\right) \bigcup \left[1-\frac{m-1}{2^L};1\right)} \left| \left(A^2_{m,2^{-L}}(g_1)_+ - A^1_{m,2^{-L}}(g_1)\right) - \left(A^2_{m,2^{-L}}(g_2)_+ - A^1_{m,2^{-L}}(g_1)\right) \right| d\lambda \\ &\leq \left\|A^1_{m,2^{-L}}(g_1) - A^1_{m,2^{-L}}(g_2)\right\|_1 + \int_{\left[0;\frac{m+1}{2^L}\right) \bigcup \left[1-\frac{m-1}{2^L};1\right)} \left|A^2_{m,2^{-L}}(g_1) - A^2_{m,2^{-L}}(g_2)\right| d\lambda. \end{aligned}$$

Now, writing P_1, P_2 the polynomials of degree m such that $A^1_{m,2^{-L}}(g_i)(t) = P_i(t)$ for $t \in \left[\frac{m}{2^L}; \frac{m+1}{2^L}\right)$. The polynomials $Q_i = P_i \circ (\cdot / \frac{m+1}{2^L})$ have degree m and, by definition, we have the two equalities,

$$\int_0^{\frac{m+1}{2^L}} \left|A^2_{m,2^{-L}}(g_1)(t) - A^2_{m,2^{-L}}(g_2)(t)\right| dt = 2^{-L}(m+1) \int_0^1 |(Q_1 - Q_2)(t)| dt,$$

$$\int_{\frac{m}{2^L}}^{\frac{m+1}{2^L}} \left|A^2_{m,2^{-L}}(g_1)(t) - A^2_{m,2^{-L}}(g_2)(t)\right| dt = 2^{-L}(m+1)\int_{\frac{m}{m+1}}^{1} |(Q_1 - Q_2)(t)|dt.$$

It remains to apply Lemma 18 to obtain

$$\begin{aligned}
&\int_0^{\frac{m+1}{2^L}} \left|A^2_{m,2^{-L}}(g_1)(t) - A^2_{m,2^{-L}}(g_2)(t)\right| dt \\
&\leq 2(m+1)^3 e^{\sqrt{6(m+1)}m} \int_{\frac{m}{2^L}}^{\frac{m+1}{2^L}} \left|A^2_{m,2^{-L}}(g_1)(t) - A^2_{m,2^{-L}}(g_2)(t)\right| dt \\
&\leq 2(m+1)^3 e^{\sqrt{6(m+1)}m} \int_{\frac{m}{2^L}}^{\frac{m+1}{2^L}} \left|A^1_{m,2^{-L}}(g_1)(t) - A^1_{m,2^{-L}}(g_2)(t)\right| dt \\
&\leq 2(m+1)^3 e^{\sqrt{6(m+1)}m} \|g_1 - g_2\|_1 .
\end{aligned}$$

Finally, we obtain the bound

$$\begin{aligned}
h(f_1, f_2) &\leq 2\frac{\left(1 + \sqrt{1 + 2(m+1)^3 e^{\sqrt{6(m+1)}m}}\right)\|g_1 - g_2\|_1^{\frac{1}{2}}}{\left\|\sqrt{A^2_{m,2^{-L}}(g_1)_+ + \tau}\right\|_2} \\
&\leq 2\left(1 + \sqrt{1 + 2(m+1)^3 e^{\sqrt{6(m+1)}m}}\right)\tau^{-1/2}\|g_1 - g_2\|_2^{\frac{1}{2}} .
\end{aligned}$$

□

Bounds on sup-norm distance.

Lemma 14. *Let $L \in \mathbb{N}$ and $(\Theta_{1,i})_{0\leq i\leq 2^L-1}$, $(\Theta_{2,i})_{0\leq i\leq 2^L-1}$ be two elements in S^{2^L}. Then, for $0 \leq m \leq 2^L - 1$ and $\epsilon > 0$, there exists a contant C depending only on m such that*

$$\begin{aligned}
&\left\|SD_{\tau,m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\Theta_{1,i}H_{Li}\right) - SD_{\tau,m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\Theta_{2,i}H_{Li}\right)\right\|_\infty \\
&\quad\leq C\left(2^L\tau^{-1} \vee 2^{2L}\tau^{-2}\right)\max_{0\leq i\leq 2^L}|\Theta_{1,i} - \Theta_{2,i}| .
\end{aligned}$$

Proof. From (2.4) and Lemma 4, it is straightforward that

$$\begin{aligned}
&\left\|A^1_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\Theta_{1,i}H_{Li}\right) - A^1_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\Theta_{2,i}H_{Li}\right)\right\|_\infty \\
&\quad\leq \left\|\sum_{0\leq i\leq 2^L-1}\Theta_{1,i}H_{Li} - \sum_{0\leq i\leq 2^L-1}\Theta_{2,i}H_{Li}\right\|_\infty \\
&\quad= 2^L\max_{0\leq i\leq 2^L-1}|\Theta_{1,i} - \Theta_{2,i}| .
\end{aligned} \tag{2.34}$$

Now, we point out that, for $J \neq \varnothing$ an interval in $[0;1)$,

$$\|f\|_{\infty,J} := \sup_{x\in J}|f(x)|$$

defines a norm on the space of polynomials of degree at most m, which is of finite dimension. Therefore there exist constants $C_{1,m} \geq 1$ and $C_{2,m} \geq 1$ such that

$$\|\cdot\|_{\infty,[0;1)} \leq C_{1,m} \|\cdot\|_{\infty,[0;(m+1)^{-1})} \text{ and } \|\cdot\|_{\infty,[0;1)} \leq C_{2,m} \|\cdot\|_{\infty,[m(m+1)^{-1};1)}.$$

Also, let's write P_j (resp. Q_j) the polynomial of degree m such that

$$A^1_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1} \Theta_{j,i} H_{Li}\right) = P_j(2^L(m+1)^{-1}, \cdot) \quad \text{resp. } Q_j\left(2^L(m+1)^{-1}\Big(\cdot -1 + (m+1)2^{-L}\Big)\right),$$

on the interval $\left[m2^{-L}; (m+1)2^{-L}\right)$ (resp. $\left[1-(m+1)2^{-L}; 1-m2^{-L}\right)$). These polynomials exist according to Lemma 8. It follows that, by definition,

$$\begin{aligned}
&\sup_{t\in[0;(m+1)2^{-L})} \left| A^2_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1} \Theta_{1,i} H_{Li}\right)(t) - A^2_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1} \Theta_{2,i} H_{Li}\right)(t)\right| \\
&\quad = \sup_{t\in[0;(m+1)2^{-L})} \left|P_1(2^L(m+1)^{-1}t) - P_2(2^L(m+1)^{-1}t)\right| \\
&\quad = \sup_{t\in[0;1)} |P_1(t) - P_2(t)| \\
&\quad \leq C_{2,m} \sup_{t\in[m(m+1)^{-1};1)} |P_1(t) - P_2(t)| \\
&\quad = C_{2,m} \sup_{t\in[m2^{-L};(m+1)2^{-L})} \left|P_1(2^L(m+1)^{-1}t) - P_2(2^L(m+1)^{-1}t)\right| \\
&\quad = C_{2,m} \sup_{t\in[m2^{-L};(m+1)2^{-L})} \left| A^2_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1} \Theta_{1,i} H_{Li}\right)(t) - A^2_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1} \Theta_{2,i} H_{Li}\right)(t)\right| \\
&\quad = C_{2,m} \sup_{t\in[m2^{-L};(m+1)2^{-L})} \left| A^1_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1} \Theta_{1,i} H_{Li}\right)(t) - A^1_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1} \Theta_{2,i} H_{Li}\right)(t)\right|.
\end{aligned}$$

Therefore, using the same arguments on $\left[1-(m+1)2^{-L}; 1\right)$ and the fact that

$$A^1_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1} \Theta_{j,i} H_{Li}\right)$$

and

$$A^2_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1} \Theta_{j,i} H_{Li}\right)$$

are equal on $\left[(m+1)2^{-L}; 1-(m+1)2^{-L}\right)$,

$$\begin{aligned}
&\left\| A^2_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1} \Theta_{1,i} H_{Li}\right) - A^2_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1} \Theta_{2,i} H_{Li}\right)\right\|_\infty \\
&\quad \leq \max(C_{1,m}, C_{2,m}) \left\| A^1_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1} \Theta_{1,i} H_{Li}\right) - A^1_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1} \Theta_{2,i} H_{Li}\right)\right\|_\infty .
\end{aligned} \tag{2.35}$$

Then, from Lemma 4 and the hypothesis on the $(\Theta_{j,i})_{0\leq i\leq 2^L-1}$ sequences, we also notice that

$$\left\| A^1_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\Theta_{j,i}H_{Li}\right)\right\|_\infty \leq 2^L.$$

Therefore, the same arguments as above gives

$$\left\| A^2_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\Theta_{j,i}H_{Li}\right)\right\|_\infty \leq 2^L \max\left(C_{1,m}, C_{2,m}\right).$$

Finally, denoting

$$I_j = \int_0^1 \left[A^2_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\Theta_{j,i}H_{Li}\right)_+ + \tau\right] d\lambda,$$

(2.8) gives

$$\begin{aligned}
&\left\| SD_{\tau,m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\Theta_{1,i}H_{Li}\right) - SD_{\tau,m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\Theta_{2,i}H_{Li}\right)\right\|_\infty \\
&\quad\leq \tau^{-1}\left\| A^2_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\Theta_{1,i}H_{Li}\right) - A^2_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\Theta_{2,i}H_{Li}\right)\right\|_\infty \\
&\qquad + \left\| A^2_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\Theta_{2,i}H_{Li}\right)\right\|_\infty \left|\frac{1}{I_1} - \frac{1}{I_2}\right| \\
&\quad\leq \left(\tau^{-1} + 2^L\max\left(C_{1,m}, C_{2,m}\right)\tau^{-2}\right) \\
&\qquad \left\| A^2_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\Theta_{1,i}H_{Li}\right) - A^2_{m,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\Theta_{2,i}H_{Li}\right)\right\|_\infty .
\end{aligned} \tag{2.36}$$

Combining (2.34), (2.35) and (2.36) concludes the proof.

□

Lemma 15. *Let $f_0 \in \Sigma(\alpha, [0,1))$, for $\alpha > 0$, be a probability density function such that $f_0 \geq \rho$ for some $\rho > 0$. Then, for $L \in \mathbb{N}$ large enough, there exists $(\eta_i)_{0\leq i\leq 2^L-1} \in S^{2^L}$ and a constant C depending only on α, $\|f_0\|_\infty$ and $\|f_0\|_{\Sigma(\alpha)}$ such that*

$$\left\| f_0 - SD_{\tau,\lfloor\alpha\rfloor,2^{-L}}\left(\sum_{i=0}^{2^L-1}\eta_i H_{Li}\right)\right\|_\infty \leq C\left(2^{-\alpha L} + \tau\right)$$

for any $\tau > 0$ small enough.

Proof. Let's write $\tilde{f}_0^L = \sum_{i=0}^{2^L-1}\eta_i S_{i,2^L,\lfloor\alpha\rfloor}$ the application from Lemma 1 such that

$$\left\| f_0\big|_{[\lfloor\alpha\rfloor 2^{-L}; 1-\lfloor\alpha\rfloor 2^{-L})} - \tilde{f}_0^L\big|_{[\lfloor\alpha\rfloor 2^{-L}; 1-\lfloor\alpha\rfloor 2^{-L})}\right\|_\infty \leq C2^{-\alpha L} \tag{2.37}$$

with $(\theta_i)_{0\leq i\leq 2^L-1} \in S^{2^L}$. Then, we see, from definitions (2.6), (2.11) and equation (2.10) that

$$\tilde{f}_0^L = A^1_{\lfloor\alpha\rfloor,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\eta_i H_{Li}\right)$$

and we introduce

$$f_0^L = SD_{\tau,\lfloor\alpha\rfloor,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\eta_i H_{Li}\right).$$

Besides, by construction (see Subsection 2.3.2) and from (2.7), there exists a polynomial P of degree $\lfloor\alpha\rfloor$ such that

$$\forall t\in\left[0;\frac{\lfloor\alpha\rfloor+1}{2^L}\right),\qquad A^2_{\lfloor\alpha\rfloor,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\eta_i H_{Li}\right)(t)=P\left(\frac{2^L t}{\lfloor\alpha\rfloor+1}\right). \tag{2.38}$$

We point out that we also have, for any $t\in\left[\frac{\lfloor\alpha\rfloor}{2^L};\frac{\lfloor\alpha\rfloor+1}{2^L}\right)$,

$$A^2_{\lfloor\alpha\rfloor,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\eta_i H_{Li}\right)(t)=A^1_{\lfloor\alpha\rfloor,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\eta_i H_{Li}\right)(t). \tag{2.39}$$

Also, if $\mathbb{R}_{\lfloor\alpha\rfloor}[X]$ is the space of all polynomials of degree at most $\lfloor\alpha\rfloor$, it is well-known fact that there exists a polynomial $Q_1\in\mathbb{R}_{\lfloor\alpha\rfloor}[X]$ such that $\forall t\in\left[0;\frac{\lfloor\alpha\rfloor+1}{2^L}\right)$, $\left|f_0(t)-Q_1(t)\right|\leq C2^{-\alpha L}$, with C a constant depending only on α and $\|f_0\|_{\Sigma(\alpha)}$. Let $Q(t)=Q_1\left(\frac{\lfloor\alpha\rfloor+1}{2^L}t\right)$ for $t\in[0;1)$. Let's define on $\mathbb{R}_{\lfloor\alpha\rfloor}[X]$, for J an interval in $[0;1)$, the norm

$$||g||_{\infty,J}:=\sup_{x\in J}|g(x)|.$$

By equivalence of norms on $\mathbb{R}_{\lfloor\alpha\rfloor}[X]$, $||g||_{\infty,[0;1)}\leq C||g||_{\infty,[\lfloor\alpha\rfloor(\lfloor\alpha\rfloor+1)^{-1};1)}$, with C a constant depending on $\lfloor\alpha\rfloor$ only.

Now, using that, by definition, $\tilde{f}_0^L$ and Q_1 are both close to f_0 on some intervals as shown above, we deduce from the last paragraph, (2.38), (2.39) and the bound (2.37)

$$\begin{aligned}
||Q-P||_{\infty,[0;1)} &= \sup_{x\in[0;1]}|Q(x)-P(x)| \\
&= \sup_{t\in\left[0;\frac{\lfloor\alpha\rfloor+1}{2^L}\right]}\left|Q\left(\frac{2^L t}{\lfloor\alpha\rfloor+1}\right)-P\left(\frac{2^L t}{\lfloor\alpha\rfloor+1}\right)\right| \\
&\leq \sup_{t\in\left[0;\frac{\lfloor\alpha\rfloor+1}{2^L}\right)}|f_0(t)-Q_1(t)|+\sup_{t\in\left[0;\frac{\lfloor\alpha\rfloor+1}{2^L}\right)}\left|f_0(t)-P\left(\frac{2^L t}{\lfloor\alpha\rfloor+1}\right)\right| \\
&\leq C2^{-\alpha L}
\end{aligned}$$

where C depends on α, $\|f_0\|_\infty$ and $\|f_0\|_{\Sigma(\alpha)}$ only. Hence, from (2.38),

$$\begin{aligned}
\sup_{t\in\left[0;\frac{\lfloor\alpha\rfloor+1}{2^L}\right)}\left|f_0(t)-A^2_{\lfloor\alpha\rfloor,2^{-L}}\left(\sum_{i=0}^{2^L-1}\theta_i 2^L\mathbb{1}_{[i2^{-L};(i+1)2^{-L})}\right)(t)\right| &\leq \sup_{t\in\left[0;\frac{\lfloor\alpha\rfloor+1}{2^L}\right)}|f_0(t)-Q_1(t)|+ \\
&\quad \sup_{t\in\left[0;\frac{\lfloor\alpha\rfloor+1}{2^L}\right)}\left|Q_1(t)-P\left(\frac{2^L t}{\lfloor\alpha\rfloor+1}\right)\right| \\
&\leq C2^{-\alpha L}+||Q-P||_{\infty,[0;1)} \\
&\leq C2^{-\alpha L}
\end{aligned}$$

with C a constant depending only on f_0, $\|f_0\|_\infty$ (which exists as f_0 is a Hölderian density, see for instance [157], p.9) and $\|f_0\|_{\Sigma(\alpha)}$. Using the same reasoning to control the distance on $\left[1-\frac{\lfloor\alpha\rfloor+1}{2^L};1\right)$ and the equality (2.39), we conclude that

$$\left\| f_0 - A^2_{\lfloor\alpha\rfloor,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\eta_i H_{Li}\right)\right\|_\infty \leq C2^{-\alpha L} \tag{2.40}$$

and $\left|1-\int_0^1 A^2_{\lfloor\alpha\rfloor,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\eta_i H_{Li}\right)(u)du\right| \leq C2^{-\alpha L}$. For L large enough, as f_0 is lower bounded by a strictly positive constant, we thus have that $A^2_{\lfloor\alpha\rfloor,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\eta_i H_{Li}\right)$ is positive for L large enough, leading to the simplification of the definition (2.8) (we remove the positive parts from the formula)

$$f_0^L = SD_{\tau,\lfloor\alpha\rfloor,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\eta_i H_{Li}\right) = \frac{A^2_{\lfloor\alpha\rfloor,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\eta_i H_{Li}\right)+\tau}{\int_0^1\left(A^2_{\lfloor\alpha\rfloor,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\eta_i H_{Li}\right)(u)+\tau\right)du}.$$

We also have that $A^2_{\lfloor\alpha\rfloor,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\eta_i H_{Li}\right)$ is upper bounded by a constant depending only on α, $\|f_0\|_\infty$ and $\|f\|_{\Sigma(\alpha)}$ as a consequence from (2.40). Finally,

$$\begin{aligned}\left\|f_0 - f_0^L\right\|_\infty \leq\ & C2^{-\alpha L}+\tau+ \\ & \left|1-\frac{1}{\int_0^1\left(A^2_{\lfloor\alpha\rfloor,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\eta_i H_{Li}\right)(u)+\tau\right)du}\right|\times \\ & \left\|A^2_{\lfloor\alpha\rfloor,2^{-L}}\left(\sum_{0\leq i\leq 2^L-1}\eta_i H_{Li}\right)(u)+\tau\right\|_\infty \\ & \leq C\left(2^{-\alpha L}+\tau\right).\end{aligned}$$

□

2.6.6 Proof of Theorem 4.

Within this proof, let us set, for c_0 to be precised,

$$\epsilon_n = c_0 n^{-\frac{\alpha}{2\alpha+1}}\log^{\frac{\alpha}{2\alpha+1}+1/2} n.$$

Let's introduce the sequences of depth $L_{1,n}$ and $L_{2,n}$ such that $2^{L_{i,n}} \asymp C_i\left(\frac{n}{\log n}\right)^{\frac{1}{2\alpha+1}}\log^{1/2} n$ for some constants C_1 and $C_2 = 1$, and introduce the subsets

$$\mathcal{F}_n = \cup_{l=1}^{L_{1,n}} G_{l,\xi(l,n),\tau_n}$$

where

$$G_{l,k,\tau_n} := \left\{ SD_{\tau_n,k,2^{-l}}(g),\ g = \sum_{i=0}^{2^l-1}\Theta_i H_{li}, (\Theta_i)_{0\leq i\leq 2^l-1}\in S^{2^l}\right\}.$$

1) Complexity of the prior: On the one hand, we have

$$\begin{aligned}\Pi[\mathcal{F}_n^c] = \Pi[l > L_{1,n}] &\lesssim 2^{-L_{1,n}^{3/2} 2^{L_{1,n}}} \\ &\leq e^{-\log^{-1/2} 2 \log^{3/2}(C_1 n/\log n) \frac{2^{L_{1,n}}}{2\alpha+1}} \\ &\leq e^{-\frac{C_1 c_0^{-2}}{(4\alpha+2)\sqrt{\log 2}} n\epsilon_n^2}. \end{aligned} \tag{2.41}$$

On the other hand, Lemma 13 implies

$$N\left(\epsilon_n, G_{l,\xi(l,n),\tau_n}, h\right) \leq N\left(C_{l,n}^{-2}\tau_n\epsilon_n^2, \left\{\sum_{i=0}^{2^l-1}\Theta_i H_{li},\ (\Theta_i)_{0\leq i\leq 2^l-1} \in S^{2^l}\right\}, h\right)$$

where $C_{l,n}$ is the multiplicative constant from Lemma 13 when $m = \xi(l,n)$. For f, g in

$$\left\{\sum_{i=0}^{2^l-1}\Theta_i H_{li}, \quad (\Theta_i)_{0\leq i\leq 2^l-1} \in S^{2^l}\right\},$$

we see that

$$h(f,g) = \left(\sum_{i=0}^{2^l-1} 2^{-l}(\sqrt{f(i2^{-l})} - \sqrt{g(i2^{-l})})^2\right)^{1/2} = \left\|\sqrt{\mathbf{f}} - \sqrt{\mathbf{g}}\right\|_2$$

where $\mathbf{f}, \mathbf{g} \in [0,1]^{2^l}$ are the sequences in S^{2^l} defining f and g. It follows that

$$N\left(\epsilon_n, G_{l,\xi(l,n),\tau_n}, h\right) \leq N\left(C_{l,n}^{-2}\tau_n\epsilon_n^2, S^{2^l}, \|\cdot\|_2\right) \leq N\left(C_{l,n}^{-2}\tau_n\epsilon_n^2, B_{\mathbb{R}^{2^l}}(0,1), \|\cdot\|_2\right) \leq \left(\frac{C}{C_{l,n}^{-2}\tau_n\epsilon_n^2}\right)^{2^l}.$$

This finally gives, since $\xi(l,n) \leq \xi(1,n) \leq \log(n)/2$, for C' an absolute constant and using the explicit formula for $C_{l,n}$ from Lemma 13,

$$\begin{aligned} N(\epsilon_n, \mathcal{F}_n, h) &\leq \sum_{l=1}^{L_{1,n}} N(\epsilon_n, G_{l,\xi(l,n),\tau_n}, h) \\ &\leq \sum_{l=1}^{L_{1,n}} \left(\frac{4C\left(1 + \sqrt{1 + 2\left(\xi(l,n)+1\right)^3 e^{\sqrt{6(\xi(l,n)+1)}\xi(l,n)}}\right)^2}{\tau_n\epsilon_n^2}\right)^{2^l} \\ &\leq \sum_{l=1}^{L_{1,n}} \left(\frac{8C\left(1 + \left(\xi(l,n)+1\right)^3 e^{\sqrt{6(\xi(l,n)+1)}\xi(l,n)}\right)}{\tau_n\epsilon_n^2}\right)^{2^l} \\ &\leq \sum_{l=1}^{L_{1,n}} \left(\frac{9Ce^{C'\log^{3/2} n}\log^3 n}{\tau_n\epsilon_n^2}\right)^{2^l} \\ &\lesssim \left(\frac{Cn^{C'\sqrt{\log n}}\log^3 n}{\tau_n\epsilon_n^2}\right)^{2^{L_{1,n}+1}} \end{aligned}$$

$$\leq n^{2C'\sqrt{\log n}2^{L_{1,n}}} \leq e^{2C'C_1c_0^{-2}n\epsilon_n^2}. \tag{2.42}$$

2) Prior mass condition: Lemma 12 ensures the existence of a sequence $(\eta_i)_{0\leq i\leq 2^{L_{2,n}}-1} \in S^{2^{L_{2,n}}}$ such that, with Θ_i the sequence drawn by the TPT distribution as in (2.1), $\max_{0\leq i\leq 2^{L_{2,n}}-1} |\eta_i - \Theta_i| \leq (\log n)^{\frac{\alpha+2}{2\alpha+1}} n^{-\frac{3\alpha+3}{2\alpha+1}}$ and

$$K\left(f_0, SD_{\tau_n,\lfloor\alpha\rfloor,2^{-L_{2,n}}}\left(\sum_{i=0}^{2^{L_{2,n}}-1}\Theta_i H_{L_{2,n}i}\right)\right) \vee V\left(f_0, SD_{\tau_n,\lfloor\alpha\rfloor,2^{-L_{2,n}}}\left(\sum_{i=0}^{2^{L_{2,n}}-1}\Theta_i H_{L_{2,n}i}\right)\right)$$

is smaller than ϵ_n^2 if c_0 is large enough, depending on ρ, α, $\|f_0\|_\infty$ and $\|f_0\|_{\Sigma(\alpha)}$. Indeed, under this condition, every term in the lemma depending on $L = L_{2,n}$ and $\tau = \sqrt{n}^{-1}$ is of the right order. Consequently,

$$\Pi\left[B_{KL}\left(f_0,\epsilon_n\right)\middle| l = L_{2,n}\right] \geq \Pi\left[\max_{0\leq i\leq 2^{L_{2,n}}-1}\left|\eta_i - \Theta_i\right| \leq \left(\log n\right)^{\frac{\alpha+2}{2\alpha+1}} n^{-\frac{3\alpha+3}{2\alpha+1}}\middle| l = L_{2,n}\right].$$

The same arguments underlying (2.17) then ensures that, for some $C > 0$, depending on β, R, α and c_0, $\Pi\left[B_{KL}\left(f_0,\epsilon_n\right)\middle| l = L_{2,n}\right] \geq e^{-Cn\epsilon_n^2}$. Therefore,

$$\begin{aligned}\Pi\left[B_{KL}(f_0,\epsilon_n)\right] &\gtrsim \Pi\left[B_{KL}(f_0,\epsilon_n)|l = L_{2,n}\right] 2^{-L_{2,n}^{3/2}2^{L_{2,n}}} \\ &\geq e^{-Cn\epsilon_n^2}e^{-\log^{3/2}n 2^{L_{2,n}}/\sqrt{\log 2}} \\ &= e^{-\left(C+\frac{1}{c_0^2\sqrt{\log 2}}\right)n\epsilon_n^2}.\end{aligned} \tag{2.43}$$

We conclude using Theorem 7 along with equations (2.41),(2.42) and (2.43) , since for C_1 large enough, $\frac{C_1c_0^{-2}}{(4\alpha+2)\sqrt{\log 2}} > C + \frac{1}{c_0^2\sqrt{\log 2}} + 4$. Then, the theorem is valid for M large enough, depending on ρ, α, $\|f_0\|_{\Sigma(\alpha)}$, $\|f_0\|_\infty$, β and R.

2.6.7 Miscellaneous.

Extension of Hölderian maps.

Lemma 16. *Let $E \subset \mathbb{R}$ be a non-empty interval and let $f_0 \in \Sigma(\alpha, E)$ where $\alpha > 0$. Then there exists a function $\widetilde{f} \in \Sigma(\alpha, \mathbb{R})$ so that $\widetilde{f}\big|_E = f_0$ and $\left\|\widetilde{f}\right\|_{\Sigma(\alpha)} = \|f_0\|_{\Sigma(\alpha)}$.*

Proof. First, let's assume that $\alpha \leq 1$. We use the fact that

$$|x_1 - x_2|^\alpha \leq (|x_1 - x_3| + |x_2 - x_3|)^\alpha \leq |x_1 - x_3|^\alpha + |x_2 - x_3|^\alpha. \tag{2.44}$$

First, for $f_0 \in \Sigma(\alpha, E)$, we have $|f_0(x) - f_0(y)| \leq \|f\|_{\Sigma(\alpha)} |x - y|^\alpha$ for all $x, y \in E$ and we define

$$h(x) := \inf\left\{f_0(y) + \|f_0\|_{\Sigma(\alpha)} |x - y|^\alpha : y \in E\right\}, \quad x \in \mathbb{R}^n.$$

If $x \in E$, then taking $y = x$ we get that $h(x) \leq f_0(x)$. To prove that $h(x)$ is finite for every $x \in \mathbb{R}^n$, fix $y_0 \in E$. If $y \in E$ then, from (2.44),

$$f_0(y) - f_0(y_0) + \|f_0\|_{\Sigma(\alpha)} |x-y|^\alpha \geq -\|f_0\|_{\Sigma(\alpha)} |y-y_0|^\alpha + \|f_0\|_{\Sigma(\alpha)} |x-y|^\alpha \geq -\|f_0\|_{\Sigma(\alpha)} |x-y_0|^\alpha,$$

and so

$$h(x) = \inf\left\{f_0(y) + \|f_0\|_{\Sigma(\alpha)} |x-y|^\alpha : y \in E\right\} \geq f_0(y_0) - \|f_0\|_{\Sigma(\alpha)} |x-y_0|^\alpha > -\infty.$$

Note that if $x \in E$, then we can choose $y_0 = x$ in the previous inequality to obtain $h(x) \geq f(x)$. Thus h extends f_0. Next we prove that

$$|h(x_1) - h(x_2)| \leq \|f_0\|_{\Sigma(\alpha)} |x_1 - x_2|^\alpha$$

for any $x_1, x_2 \in \mathbb{R}$. Given $\varepsilon > 0$, by the definition of h there exists $y_1 \in E$ such that

$$h(x_1) \geq f_0(y_1) + \|f_0\|_{\Sigma(\alpha)} |x_1 - y_1|^\alpha - \varepsilon.$$

Since $h(x_2) \leq f_0(y_1) + \|f_0\|_{\Sigma(\alpha)} |x_2 - y_1|^\alpha$, we get from (2.44)

$$\begin{aligned} h(x_1) - h(x_2) &\geq \|f_0\|_{\Sigma(\alpha)} |x_1 - y_1|^\alpha - \|f_0\|_{\Sigma(\alpha)} |x_2 - y_1|^\alpha - \varepsilon \\ &\geq -\|f_0\|_{\Sigma(\alpha)} |x_1 - x_2|^\alpha - \varepsilon. \end{aligned}$$

Letting $\varepsilon \to 0$ gives $h(x_1) - h(x_2) \geq -\|f_0\|_{\Sigma(\alpha)} |x_1 - x_2|^\alpha$. It remains to reverse the roles of x_1 and x_2 to prove that h is Hölder continuous with $\|h\|_{\Sigma(\alpha)} = \|f_0\|_{\Sigma(\alpha)}$.
Now, if $\alpha > 1$, we have that $f_0 \in \Sigma(\alpha, E)$ implies

$$f_0^{(\lfloor\alpha\rfloor)} \in \Sigma(\alpha - \lfloor\alpha\rfloor, E), \quad \left\|f_0^{(\lfloor\alpha\rfloor)}\right\|_{\Sigma(\alpha-\lfloor\alpha\rfloor)} = \|f_0\|_{\Sigma(\alpha)}.$$

The above proof ensures that there exist $g \in \Sigma(\alpha - \lfloor\alpha\rfloor, \mathbb{R})$ such that $g\big|_E = f_0^{(\lfloor\alpha\rfloor)}$. As this last application as well as g are continuous, it suffices to take the successive primitives of g (with equality constraint ensuring that these are also derivatives of f_0) to obtain the result. □

Control of discretization error.

Lemma 17. *Let f be a piecewise constant map on the intervals $[i/l; (i+1)/l)$, $i = 0, \ldots, l-1$, with $l \in \mathbb{N}^*$. Let's assume that f takes values in $[0; l']$. Then, if g is the 1-periodic extension of f on $\mathbb{R}$, we have*

$$\left\|g_{\infty,s}^m - g_{q,s}^m\right\|_\infty \leq \frac{ml'(ls+1)}{q}$$

for $m \in \mathbb{N}$, $0 < s < 1/2$.

Proof. If $m = 0$, the result is straightforward. Otherwise, we start by noting that, as g is 1-periodic, (2.3) gives that $g_{\infty,s}^m$, as well as $g_{q,s}^m$, are themselves 1-periodic. It is therefore sufficient to control $\left|g_{\infty,s}^m(x) - g_{q,s}^m(x)\right|$ for x in $[0; 1)$.
Before going further, let's first show that for any $m \geq 0$, $g_{\infty,s}^m$ is a function of bounded variation over any interval $[x-s; x]$, with a bound on its total variation independent of $x \in \mathbb{R}$. This means that there exists a constant $V = V(l', l, s) > 0$, such that, for any $x \in \mathbb{R}$ and subdivision $\sigma = \{x - s = x_1 < x_2 \ldots x_{n-1} < x_n = x \,|\, n \geq 2\}$,

$$\sum_{i=1}^{n} \left|g_{\infty,s}^m(x_{i+1}) - g_{\infty,s}^m(x_i)\right| \leq V.$$

By assumptions on f, on an interval of length s, g is piecewise constant with at most $ls+1$ discontinuity points and takes values in $[0; l']$. Therefore, g has bounded variation at most

$V = l'(ls+1)$. Then we show that the convolution of $s^{-m}\chi_s^{*m}$ with g is also a function with bounded variation on any interval $[x-s;x]$ with $x \in [0;1)$. Indeed, for a subdivision σ of $[x-s;x]$ and m such that $s(m+1) < 1$, using (2.3) and Lemma 4

$$\begin{aligned}\sum_{i=1}^{n} \left|g_{\infty,s}^m(x_{i+1}) - g_{\infty,s}^m(x_i)\right| &= \sum_{i=1}^{n} \left|s^{-m}\chi_s^{*m} * g(x_{i+1}) - s^{-m}\chi_s^{*m} * g(x_i)\right| \\ &\leq \int_{\mathbb{R}} \left[\sum_{i=1}^{n} \left|g(x_{i+1}-u) - g(x_i-u)\right|\right] \left|s^{-m}\chi_s^{*m}(u)\right| \, du \\ &\leq l'(ls+1) \int_{\mathbb{R}} s^{-m}\chi_s^{*m}(u) \, du \qquad (2.45) \\ &= V.\end{aligned}$$

Let's assume $m = 1$, in which case $g_{q,s}^1(x)$ is a Riemann sum with converges to $g_{\infty,s}^1(x) = s^{-1}\int_{x-s}^{x} g(t)dt$ as $q \to \infty$. More precisely, with the bounded variation property above, it is a common result that in this case

$$\left|sg_{\infty,s}^1(x) - sg_{q,s}^1(x)\right| \leq V\frac{s}{q} \leq \frac{l's(ls+1)}{q}.$$

This proves the lemma for $m = 1$.
For the general case, let's assume that the lemma is true for some $m \in \mathbb{N}^*$. We use the same argument, since the above equation translates in

$$\left|g_{\infty,s}^{m+1}(x) - \left(g_{\infty,s}^m\right)_{q,s}^1(x)\right| \leq \frac{l'(ls+1)}{q},$$

according to (2.45). Also, the property at level m ensures that

$$\begin{aligned}\left|g_{q,s}^{m+1}(x) - \left(g_{\infty,s}^m\right)_{q,s}^1(x)\right| &\leq \frac{1}{q}\sum_{i=0}^{q-1}\left|g_{q,s}^m\left(x - \frac{is}{q}\right) - g_{\infty,s}^m\left(x - \frac{is}{q}\right)\right| \\ &\leq \frac{ml'(ls+1)}{q}.\end{aligned}$$

Using the triangular inequality on $\left|g_{\infty,s}^{m+1}(x) - g_{q,s}^{m+1}(x)\right|$ to obtain a bound from the sum of the two terms above allows to conclude the proof. □

Equivalence of norms on spaces of polynomials.

Lemma 18. *Let $\mathbb{R}_n[X]$ be the space of polynomials of degree at most n and with real coefficients. For $P \in \mathbb{R}_n[X]$, let's write*

$$\|P\|_{\infty,J} \coloneqq \max_{t\in J}|P(t)|, \quad \textit{with } J \textit{ an interval in } \mathbb{R}.$$

Then, for all $P \in \mathbb{R}_n[X]$ and $s \in [0;1)$, we have

$$\|P\|_{\infty,[0;1]} \leq e^{\sqrt{6s^{-1}}n} \|P\|_{\infty,[0;s]}$$

as well as

$$\|P\|_{\infty,[0;1]} \leq e^{\sqrt{6s^{-1}}n} \|P\|_{\infty,[1-s;1]}.$$

It also follows that

$$\int_{[0;1]} |P(u)|\,du \leq e^{\sqrt{6s^{-1}}n}\frac{2}{s}(n+1)^2\int_{[0;s]} |P(u)|\,du,$$

$$\int_{[0;1]} |P(u)|\,du \leq e^{\sqrt{6s^{-1}}n}\frac{2}{s}(n+1)^2\int_{[1-s;1]} |P(u)|\,du.$$

Proof. For $n=0$, the result is straightforward. Let's then delve into the case $n\geq 1$. First, as

$$x^k - y^k = (x-y)\sum_{i=0}^{k-1} x^i y^{k-1-i}$$

for any x, y in $[-1;1]$,

$$|x^k - y^k| \leq k|x-y|.$$

Then, if $P(x)=\sum_{k=0}^n a_k x^k$, we have

$$|P(x)-P(y)| \leq \sum_{k=1}^n |a_k||x^k-y^k| \leq |x-y|\sum_{k=1}^n k|a_k|.$$

Before going further, we point out that necessarily $a_k = P^{(k)}(0)/(k!)$. Now, let's recall Markov's brother inequality (that can be found in [18] for instance) which states that, for $Q\in\mathbb{R}_n[X]$ and any nonnegative integers k

$$\max_{-1\leq x\leq 1}|Q^{(k)}(x)| \leq \frac{n^2(n^2-1^2)(n^2-2^2)\cdots(n^2-(k-1)^2)}{1\cdot 3\cdot 5\cdots(2k-1)}\max_{-1\leq x\leq 1}|Q(x)|.$$

The constant appearing in the above inequality is equal to $T_n^{(2k)}(1)$, where T_n is the n-th Chebyshev polynomial. Let's apply this inequality to $Q = P\circ\left(\frac{s}{2}(X+1)\right)$. First, we have

$$\|P\|_{\infty,[0;s]} = \max_{-1\leq x\leq 1}|Q(x)|.$$

And then, for $x\in[-1;1]$,

$$Q^{(k)}(x) = \left(\frac{s}{2}\right)^k P^{(k)}\left(\frac{s}{2}(x+1)\right),$$

so that

$$\max_{-1\leq x\leq 1}|Q^{(k)}(x)| = \left(\frac{s}{2}\right)^k \max_{0\leq x\leq s}|P^{(k)}(x)| \geq \left(\frac{s}{2}\right)^k |P^{(k)}(0)|.$$

Combining these results finally gives us, for $0\leq x,y\leq 1$,

$$\begin{aligned}
|P(x)-P(y)| &\leq |x-y|\sum_{k=1}^n k|a_k| \\
&= |x-y|\sum_{k=1}^n \frac{|P^{(k)}(0)|}{(k-1)!} \\
&\leq |x-y|\sum_{k=1}^n \left(\frac{2}{s}\right)^k \frac{T_n^{(2k)}(1)\,\|P\|_{\infty,[0;s]}}{(k-1)!}.
\end{aligned}$$

It follows readily that

$$\|P\|_{\infty,[0;1]} \leq \left(1 + \sum_{k=1}^{n} \left(\frac{2}{s}\right)^k \frac{T_n^{(2k)}(1)}{(k-1)!}\right) \|P\|_{\infty,[0;s]} .$$

The multiplicative factor above is then bounded by

$$\begin{aligned} 1 + \sum_{k=1}^{n} \frac{6^k s^{-k} n^{2k}}{(2k)!} &\leq 1 + \sum_{k=1}^{n} \frac{(\sqrt{6s^{-1}}n)^{2k}}{(2k)!} \\ &\leq e^{\sqrt{6s^{-1}}n} \end{aligned}$$

as $k \leq (3/2)^k, \forall k \geq 1$. This concludes the proof of the first inequality. For the second inequality, it suffices to remark that for $P \in \mathbb{R}_n[X]$,

$$\|P\|_{\infty,[1-s;1]} = \|R\|_{\infty,[0;s]} , \qquad \|P\|_{\infty,[0;1]} = \|R\|_{\infty,[0;1]} ,$$

with $R(t) = P(1-t)$ for any t a real number, defining R as an element of $\mathbb{R}_n[X]$.

Finally, for the last claim, we introduce the primitive $p(\cdot) = \int_0^{\cdot} P(t)dt$ which is a polynomial of degree at most $n+1$ verifying

$$\|p\|_{\infty,[0;s]} \leq \int_0^s |P(t)|dt.$$

With Markov's brother inequality and rescaling, we have

$$\|P\|_{\infty,[0;s]} \leq \frac{2}{s}(n+1)^2 \|p\|_{\infty,[0;s]}$$

which allows us to conclude

$$\int_{[0;1]} |P(u)|\, du \leq \|P\|_{\infty,[0;1)} \leq e^{\sqrt{6s^{-1}}n} \|P\|_{\infty,[0;s]} \leq e^{\sqrt{6s^{-1}}n} \frac{2}{s}(n+1)^2 \int_0^s |P(u)|du.$$

□

2.6.8 Random shifts for the Pólya forest

As for the question of the addition of random shifts in the DPA prior, it appears that our proof for the posterior contraction rate doesn't extend well. It mainly pertains to the difficulty of defining a low-dimensional sieve on which the prior distribution concentrates its mass. However, the control of prior mass on KL-balls centered on the true density f_0 can be conducted even with random shifts in the definition of trees in the prior. This naturally leads to a study of $\rho-$posterior contraction rates, i.e. for some $\rho \geq 0$, a prior Π on densities f and observations $X^{(n)}$ from $\mathbb{P}_{f_0}^{\otimes n}$ giving the posterior distribution Π_n,

$$\Pi_n^{(\rho)}\left[f \in B\right] = \frac{\int_B \prod_{i=1}^n f^\rho(X_i) d\Pi(p)}{\int \prod_{i=1}^n f^\rho(X_i) d\Pi(p)}$$

Indeed, according to Theorem 8.43 and Example 8.44 in [65], it suffices to show that $\Pi\left[f : K(f_0, f) < \epsilon_n^2\right] \geq \exp(-Cn\epsilon_n^2)$ for some $C \geq 0$ to obtain that the $\rho-$posterior concentrates around f_0 at rate ϵ_n,

$$\Pi_n^{(\rho)}\left[h(f_0, f) > M_n \epsilon_n\right] \to 0,$$

for any $M_n \to \infty$ and h the Hellinger distance, whenever $\rho \in (0,1)$. This statement is true for any sequence ϵ_n, but we use $r > 0$ and

$$\epsilon_n(r) = \left(\frac{\log n}{n}\right)^{\frac{r}{2r+1}},$$

as we assume $f_0 \in \Sigma(r, [0;1))$, i.e., it has Hölder regularity of degree r.

New prior RDPA.

In this subsection, we define a new prior, similar to DPA, with the exception that the shifts of the partitions underlying individual trees are random. DPA outputs maps of the form $f^m_{q,2^{-L}}$, with $q > 0$, $L > 0$, $m \geq 0$ and $f = \sum_{i\in\mathbb{Z}} u_i H_{Li}$. Instead, RDPA outputs a stochastic transform of f as is made explicit in the below algorithm defining this new distribution.

1. Fix $L > 0$, parameters set $\mathcal{A}$, $m \geq 0$, $q > 0$.
2. Draw g such that $g \sim TPT_L(\mathcal{A})$. One writes
$$g(\cdot) = \sum_{i=0}^{2^L-1} \Theta_i H_{Li}(\cdot)$$
for some sequence $(\Theta_i)_{i=0}^{2^L-1}$ whose elements are positive and sum up to 1.
3. If $m \neq 0$, draw independent random variables $U_{j,i,d} \sim \mathcal{U}[0;1]$, $1 \leq j \leq m$, $1 \leq i \leq q^{j-1}$, $1 \leq d \leq q$.
4. If $m \neq 0$, outputs
$$f_{q,m}(\cdot) = \left(\sum_{1\leq j\leq m} q^j\right)^{-1} \sum_{(i_1,\dots,i_m),\ i_j\in[\![1;q^{j-1}]\!]} \sum_{(d_1,\dots,d_m)\in[\![1;q]\!]^m} g\left(\cdot - \frac{U_{1,i_1,d_1} + \dots + U_{m,i_m,d_m}}{2^L}\right), \tag{2.46}$$
otherwise outputs $f_{q,0} = g$.

Then, we refer to the law of $f_{q,m} \,\Big|\, L, \mathcal{A}, m, q$ as $RDPA(L, \mathcal{A}, m, q)$.

N.B.: Above and in the next subsection, the difference in the arguments of maps are to be considered congruent modulo 1.

$\rho-$posterior contraction rate.

Theorem 8. *Suppose $f_0 \in \Sigma(r, [0,1))$, $r > 0$ and $f_0 \geq \rho$ for some $\rho > 0$. Let us endow f with a $RDPA(L_n, \mathcal{A}, m, q_n)$ prior which we write Π_n, such that for some $\beta > 0, R \geq 1, \delta > 0$,*

$$\forall a \in \mathcal{A},\ a \in \left[\delta\left(\frac{\log n}{n}\right)^{\beta}; R\right],$$

$L_n \asymp (n/\log n)^{1/(2r+1)}$, $m = \lfloor r \rfloor$ and $q_n^m = Cn^{2r/(2r+1)} \log^{2r/(2r+1)} n$ for C large enough. Then, for any $M_n \to \infty$ and h the Hellinger distance, as $n \to \infty$,

$$\Pi_n^{(\rho)}\left[h(f_0, f) > M_n \epsilon_n(r)\right] \to 0,$$

whenever $\rho \in (0,1)$.

Proof. As stated above, it suffices to control the prior mass of the ball $\{K(f_0, f) < \epsilon_n(r)^2\}$. In the following, we simply note $\epsilon_n = \epsilon_n(r)$.
According to the proof of Theorem 2, there exists an event $\mathcal{B}_n$, $\Pi[\mathcal{B}_n] \geq \exp(-C_1 n\epsilon_n^2)$, on which $(\Theta)_{i=0,\cdots,2^{L_n}}$ is such that

$$K\left(f_0, g^m_{\infty,2^{-L_n}}\right) < c\epsilon_n^2$$

with $c > 0$ a small constant and $2^{L_n}\Theta_i \leq C_{f_0}$ for any $i = 0, \cdots, 2^{L_n}$, where C_{f_0} is a constant depending on the sup-norm and the Hölder norm of f_0. Also, replacing

$$\frac{g^m_{q-n,2^{-L_n}}}{\int_0^1 g^m_{q_n,2^{-L_n}}(t)dt}$$

by $f_{q_n,m}$ in the proof of Lemma 10, we prove that on $\mathcal{B}_n$, if $\left\|f_{q_n,m} - g^m_{\infty,2^{-L_n}}\right\|_\infty < c'\epsilon_n$ for $c' > 0$ small enough, then

$$K\left(f_0, f_{q_n,m}\right) < \epsilon_n^2.$$

Consequently,

$$\Pi\left[f : \ K(f_0, f) < \epsilon_n^2\right] \geq \Pi[\mathcal{B}_n]\,\Pi\left[\left\|f_{q_n,m} - g^m_{\infty,2^{-L_n}}\right\|_\infty < c'\epsilon_n \middle| \mathcal{B}_n\right] \tag{2.47}$$

$$\geq \exp(-C_1 n\epsilon_n^2)\Pi\left[\left\|f_{q_n,m} - g^m_{\infty,2^{-L_n}}\right\|_\infty < c'\epsilon_n \middle| \mathcal{B}_n\right]. \tag{2.48}$$

Since the law of $U_{1,i_1,d_1} + \cdots + U_{m,i_m,d_m}$ is the one with density χ^{*m}, for any $x \in [0;1)$, $f_{q,m}(x)$ has expectation

$$\mathcal{E}_{t\sim\chi^{*m}}\left[g(x - t2^{-L_n})\right] = \int_{\mathbb{R}} g\left(x - t2^{-L_n}\right)\chi^{*m}(t)dt = 2^{-L_n}\chi^{*m}(2^{-L_n}\cdot) * g(x) = g^m_{\infty,2^{-L_n}}(x).$$

Then, $f_{q_n,m}(x)$ is the expectation of the same quantity as above w.r.t. to the empirical measure $\mathcal{E}_{q_n,m}$ associated to $\sum_{1\leq j\leq m} q_n^j \asymp q_n^m$ independent random variables with probability density χ^{*m}. Consequently, for the family of maps $\mathcal{F} = \left\{g(x - 2^{-L_n}\cdot),\ x \in [0;1)\right\}$,

$$\left\|f_{q_n,m} - g^m_{\infty,2^{-L_n}}\right\|_\infty = \sup_{f\in\mathcal{F}} |\mathcal{E}_{t\sim\chi^{*m}}[f] - \mathcal{E}_{q_n,m}[f]| =: \|e_{q_n,m}\|_{\mathcal{F}},$$

where $e_n = \mathcal{E}_{t\sim\chi^{*m}} - \mathcal{E}_{q_n,m}$ above is the empirical process indexed by $\mathcal{F}$.
Now, we would like to upper bound $\mathcal{E}\left[\|e_n\|_{\mathcal{F}}\right]$ by mean of Theorem 10. It is possible to replace $\mathcal{F}$ by $0 \bigcup \mathcal{F}$ as it only increases the expectation of the supremum. An enveloppe F such that $|f| \leq F, \forall f \in \mathcal{F}$ is $F(\cdot) = 2^{L_n}\sup_{i=0,\ldots,2^{L_n}-1}\Theta_i$. Then, for any $f = g(x - 2^{-L_n}\cdot) \in \mathcal{F}$, we have like above that

$$\mathcal{E}_{t\sim\chi^{*m}}\left[f^2(t)\right] = \mathcal{E}_{t\sim\chi^{*m}}\left[g^2(x - t2^{-L_n})\right] = 2^{-L_n}\chi^{*m}(2^{-L_n}\cdot) * g^2(x).$$

Therefore, using the fact that the $S_{i,2^{L_n},m}$ do a partition of the unity, up to a constant,

$$\sigma_n^2 := \sup_{f\in\mathcal{F}} \mathbb{E}\left[f^2(X)\right] = \left\|\sum_{i=0}^{2^{L_n}-1} \Theta_i^2 S_{i,2^{L_n},m}\right\|_\infty \gtrsim 2^{L_n} \sup_{i=0,\ldots,2^L=1} \Theta_i^2,$$

where the last inequality comes from Lemma 5 in [65]. This Lemma also proves that $\sigma_n \leq 2^{L_n/2} \sup_{i=0,\dots,2^L=1} \Theta_i \leq C_{f_0}$. Before concluding, we note that that for $\tau > 0$, Lemma 19 ensures

$$\log\left(2N(\mathcal{F}, L^2(Q), \tau \|F\|_{L^2(Q)})\right) \leq H(\tau^{-1}) = \begin{cases} \log\left(\frac{2^{L_n+1}}{\tau}\right), & \text{if } \tau^{-1} > 1 \\ \log 2, & \text{if } \tau^{-1} \leq 1 \end{cases}$$

for discrete measures Q with finite number of atoms and rational weights and N the covering numbering of a set. Then, according to Lemma 20 and Theorem 10 with $U = C_{f_0}$, for n large enough and some constant C depending on the sup-norm and the Hölder norm of f_0,

$$\mathcal{E}\left[\|e_{q_n,m}\|_{\mathcal{F}}\right] \leq C\frac{\sqrt{H\left(2^{L_n/2}\right)}}{\sqrt{\sum_{1\leq j\leq m} q_n^j}} \leq C q_n^{-m/2}\sqrt{\log(22^{3L_n/2})} \leq c'\epsilon_n/2.$$

Now, this implies, along with Theorem 9 applied to $\{f - \mathcal{E}_{t\sim\chi^{*m}}[f] : \ f \in \mathcal{F}\}$ and the $U_{1,i_1,d_1} + \cdots + U_{m,i_m,d_m}$ random variables, with $U = F$ defined above,

$$\begin{aligned}\Pi\left[\|e_{q_n,m}\|_{\mathcal{F}} \geq c'\epsilon_n \middle| \mathcal{B}_n\right] &\leq \Pi\left[\|e_{q_n,m}\|_{\mathcal{F}} \geq \mathcal{E}\left[\|e_{q_n,m}\|_{\mathcal{F}}\right] + c'\epsilon_n/2 \middle| \mathcal{B}_n\right] \\ &\leq \exp\left(-\frac{q_n^{2m}\epsilon_n^2}{4F\mathcal{E}\left[\|e_{q_n,m}\|_{\mathcal{F}}\right] + 2q_n^m F^2 + 2q_n^m\epsilon_n F/3}\right) \\ &\leq \exp\left(-Cq_n^m\epsilon_n^2\right) \to 0.\end{aligned}$$

We conclude with 2.47. □

Lemma 19. *For $L \geq 0$, $(a_k)_{k=0,\dots,2^L-1} \in S^{2^L}$, let*

$$g = \sum_{k=0,\dots,2^L-1} a_k 2^L \mathbb{1}_{[k2^{-L},(k+1)2^{-L})}.$$

Then, for $\mathcal{F} = \left\{g(x - 2^{-L}\cdot),\ x \in [0;1)\right\}$ and $F(\cdot) = 2^L \sup_k a_k$,

$$N(\mathcal{F}, L^2(Q), \tau\|F\|_{L^2(Q)}) \leq \begin{cases} \frac{2^L}{\tau}, & \text{if } \tau < 1 \\ 1, & \text{if } \tau \geq 1 \end{cases}$$

for any discrete measure Q on $\mathbb{R}$ with finite number of atoms and rational weights.

Proof. Given the assumptions on the a_k's, it is clear that the set $\{f\}$ covers $\mathcal{F}$ for any measure Q whenever $\tau \geq 1$. Otherwise, if $\tau < 1$, let's define $\nu_i = \inf\left\{\nu \in [0;1] : Q\left(\bigcup_{n\in\mathbb{Z}}[n; n+\nu]\right) \geq i\tau\right\}$ for $i = 0, \dots, \lfloor\tau^{-1}\rfloor$. One notes that $\forall i, \nu_i < 1$ as Q has a finite number of atoms. Then for $x \in [0;1)]$, take i such that $\nu_i = \sup\{\nu_j : \ 2^{-L}(\lfloor 2^L x\rfloor + \nu_j) \leq x\}$ and $y = 2^{-L}(\lfloor 2^L x\rfloor + \nu_i)$. Then, with $a_{2^L-1} = a_{-1}$ and $\nu_{\lfloor\tau^{-1}\rfloor+1} = 1$,

$$\begin{aligned}\left\|g(x - 2^{-L}\cdot) - g(y - 2^{-L}\cdot)\right\|^2_{L^2(Q)} &\leq \sum_{0\leq k\leq 2^L-1} |a_k - a_{k-1}|^2 2^{2L} Q\left((2^L y - k; 2^L x - k]\right) \\ &\leq F^2 Q\left(\bigcup_{0\leq k\leq 2^L-1} (2^L y - k; 2^L x - k]\right)\end{aligned}$$

$$
\begin{aligned}
&= \|F\|^2_{L^2(Q)} Q\left(\bigcup_{0\le k\le 2^L-1} (2^L y-k; 2^L x-k]\right)\\
&\le \|F\|^2_{L^2(Q)} Q\left(\bigcup_{0\le k\le 2^L-1} (2^L y-k; 2^L x-k]\right)\\
&\le \|F\|^2_{L^2(Q)} Q\left(\bigcup_{n\in\mathbb{Z}} (n+\nu_i; n+2^L x-\lfloor 2^L x\rfloor]\right)\\
&\le \|F\|^2_{L^2(Q)} \left[Q\left(\bigcup_{n\in\mathbb{Z}} [n; n+\nu_{i+1})\right) - Q\left(\bigcup_{n\in\mathbb{Z}} [n; n+\nu_i]\right)\right]\\
&\le \tau \|F\|^2_{L^2(Q)}\,.
\end{aligned}
$$

The last inequality comes from the fact that since Q is discrete with finite number of atoms, there exists $\nu < \nu_{i+1}$ such that $Q\left(\bigcup_{n\in\mathbb{Z}}[n; n+\nu_{i+1})\right) = Q\left(\bigcup_{n\in\mathbb{Z}}[n; n+\nu]\right)$, and by definition, $Q\left(\bigcup_{n\in\mathbb{Z}}[n; n+\nu]\right) \le (i+1)\tau \wedge 1$. Also, by continuity, $Q\left(\bigcup_{n\in\mathbb{Z}}[n; n+\nu_i]\right) \ge i\tau$. Therefore,

$$
\tilde{\mathcal{F}} = \left\{g(2^{-L}(m+\nu_i) - 2^{-L}\cdot),\ m = 0,\dots,2^L-1,\ i = 0,\dots,\lfloor \tau^{-1}\rfloor\right\}
$$

is a covering of size $2^L\lfloor\tau^{-1}\rfloor \le 2^L\tau^{-1}$. □

Lemma 20. *For any $A > 2$, the application*

$$
H : x \mapsto \begin{cases} \log 2, & \text{if } 0 < x \le 1 \\ \log(Ax), & \text{if } x > 1 \end{cases}
$$

satisifies the conditions of Theorem 10 with $C_H = 1$.

Proof. The two first points of Theorem 10 are clear. When it comes to the third one, for $0 < c \le 1$,

$$
\begin{aligned}
\int_0^c \sqrt{H(1/x)}dx &= \int_0^c \sqrt{\log(A/x)}dx\\
&= \int_{\sqrt{\log\left(\frac{A}{c}\right)}}^{+\infty} 2Au^2 e^{-u^2} du \qquad \text{with change of variable } u = \sqrt{\log\left(\frac{A}{x}\right)}\\
&\le A\sqrt{\log\left(\frac{A}{c}\right)} \int_{\sqrt{\log\left(\frac{A}{c}\right)}}^{+\infty} 2ue^{-u^2} du\\
&= c\sqrt{\log\left(\frac{A}{c}\right)}.
\end{aligned}
$$

□

Useful results.

Theorem 9 (Talagrand's inequality, Theorem 3.3.9 of [69, p.156])**.** *Let $(S, \mathcal{S})$ be a measurable space, and let $n \in \mathbb{N}$. Let $(X_1, \dots, X_n)$ be independent $S-$valued random variables. For*

$\mathcal{F}$ a countable set of measurable real-valued functions on S such that $\|f\|_\infty \leq U < \infty$ and $\mathbb{E}[f(X_1)] = \cdots = \mathbb{E}[f(X_n)] = 0$ for all $f \in \mathcal{F}$, let

$$S_n = \sup_{f \in \mathcal{F}} \left| \sum_{k=1}^{n} f(X_k) \right|$$

and parameters σ_n^2 and ν_n be defined by

$$\frac{1}{n} \sum_{k=1}^{n} \sup_{f \in \mathcal{F}} \mathbb{E}\left[f^2(X_k)\right] \leq \sigma_n^2 \leq U^2$$

and

$$\nu_n = 2U\mathbb{E}[S_n] + n\sigma_n^2.$$

Then, for any $x \geq 0$,

$$\mathbb{P}[S_n \geq \mathbb{E}[S_n] + x] \leq \exp\left(-\frac{x^2}{2\nu_n + 2xU/3}\right).$$

In the preceding subsection, we have interest in the empirical process indexed by a set of functions $\mathcal{F}$ defined by

$$f \mapsto e_n(f) := \sqrt{n}\,(Pf - P_n f)$$

where Qf denotes the expectation operator under the distribution Q, P is a probability measure and $P_n := n^{-1} \sum_{i=1}^{n} \delta_{X_i}$ is the empirical measure associated to $(X_1, \ldots, X_n) \sim P^{\otimes n}$. Also, in order to apply the preceding theorem to this setting, one needs to bound

$$\|e_n\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |e_n(f)|.$$

Theorem 10 (Theorem 3.5.6 of [69, p.189]). *Let $\mathcal{F}$ be a countable class of functions with $0 \in \mathcal{F}$ and F measurable be such that $|f| \leq F$ for all $f \in \mathcal{F}$. Also, let's write $\sigma^2 = \sup_{f \in \mathcal{F}} \mathbb{E}[f^2(X)]$ and $U = \max_{1 \leq i \leq n} F(X_i)$. Then, let $H : \mathbb{R}^+ \mapsto \mathbb{R}^+$ be a function such that*

- *$H(x) = \log 2$ for $0 < x \leq 1$;*
- *H is nondecreasing for $x > 0$, and so is $x \mapsto xH^{1/2}(x)$ for $0 < x \leq 1$;*
- *There exists $0 < C_h < +\infty$ such that $\int_0^c \sqrt{H(1/x)}dx \leq C_H c H^{1/2}(1/c)$ for all $0 < c \leq 1$.*

Assume that

$$\sup_Q \log\left(2N(\mathcal{F}, L^2(Q), \tau \|F\|_{L^2(Q)})\right) \leq H\left(\frac{1}{\tau}\right), \quad \text{for all } \tau > 0,$$

where the supremum is over all discrete measures Q with finite number of atoms and rational weights. Then,

$$\mathbb{E}\left[\left\|\sqrt{n}\,(P - P_n)\right\|_{\mathcal{F}}\right] \leq \max\left[8\sqrt{6}C_H\sigma\sqrt{H\left(\|F\|_{L^2(P)}/\sigma\right)}, \right.$$
$$\left. 2^{15}3^{5/2}C_H^2 \|U\|_{L^2(P)} \frac{H\left(\|F\|_{L^2(P)}/\sigma\right)}{\sqrt{n}}\right].$$

CHAPTER 3

Optional Pólya Trees

We consider statistical inference in the density estimation model using a tree–based Bayesian approach, with Optional Pólya trees as prior distribution. We derive near-optimal convergence rates for corresponding posterior distributions with respect to the supremum norm. For broad classes of Hölder–smooth densities, we show that the method automatically adapts to the unknown Hölder regularity parameter. We consider the question of uncertainty quantification by providing mathematical guarantees for credible sets from the obtained posterior distributions, leading to near–optimal uncertainty quantification for the density function, as well as related functionals such as the cumulative distribution function. The results are illustrated through a brief simulation study.

Table of Contents

3.1 Introduction

Tree–based methods are among the most broadly used algorithms in statistics and machine learning. This goes from single tree algorithms such as CART [20] or Bayesian CART [39, 49], to the use of random forests [13, 38], that is ensembles of trees. Due in particular to their ability to quantify uncertainty, there has been much interest in Bayesian tree–based methods. While for frequentist methods there is a by now well–established theory in quadratic loss for CART and related algorithms, advances on the mathematical understanding of Bayesian counterparts are very recent. In [145, 107], L^2–posterior contraction rates are obtained for both trees and forests in a regression setting. Still in regression, the work [34] addresses the case of the stronger supremum norm loss for Bayesian CART–type priors. The present chapter can be seen as a continuation of [34], investigating the density estimation setting. In Bayesian density estimation, a classical tree–method is that of Pólya trees (henceforth PTs, see e.g. [65], Chapter 3). For well–chosen parameters, PTs' samples are random densities, and contraction rates for the corresponding posterior densities have been obtained in [29]. The idea behind Pólya tree is to grow a fixed, infinite, tree; this is typically not flexible enough to address refined statistical goals such as adaptation. Notably, Wong and Ma introduced in [173] a flexible alternative to standard PTs that they call Optional Pólya Trees (OPTs in the sequel), which have been successfully extended and applied to a number of settings in e.g. [114, 83, 113, 110, 42]. Yet, from the theoretical point of view, only posterior consistency was established in [173] and follow-up works. Not based on (flexible) trees, we also note the different construction of spike–and–slab Pólya trees introduced in [31].

There are two main goals in the present chapter. The first is to continue the investigations of [34] for tree–methods in order to obtain inference in the practically very desirable supremum norm loss, but in the model of density estimation, and the second to elaborate a theory for rates and uncertainty quantification (henceforth, UQ) for Optional Pólya Trees. In fact, our methods enable to cover also more general priors, although for simplicity we will mostly stick to OPTs in this work. We now briefly review a number of related results. While the use of a general theory based on prior mass and testing [64, 65] made a relatively broad L^2–theory possible [107, 145], results for the supremum norm are typically more delicate, as uniform testing rates required in [64] appear to be slower [68]. Recent advances on this front include [28, 78, 125, 122, 177]. The first supremum norm posterior rates for tree methods, optimal up to a logarithmic factor, were obtained in [34] in regression models; we refer to [34] for more context and references on rates for tree–based methods.

The main results of this chapter are as follows

1. we prove that Optional Pólya Trees (OPTs) achieve optimal supremum–norm posterior

contraction rates (up to a logarithmic factor) in density estimation: this provides an optimal rate–theory for the consistency results of [173], who introduced the OPT prior, for the computationally efficient case of dyadic splits.

2. we show that tree–based inference with OPTs leads to (near–) optimal uncertainty quantification in terms of confidence bands, both for the density f and the distribution function $F = \int_0^{\cdot} f$, in an adaptive way.

Those constitute the first results, to the best of our knowledge, showing that tree–based methods in density estimation lead to near–optimal uncertainty quantification in terms of the supremum norm. Apart from making the consistency results of [173] precise, this work shows that the programme for inference with tree–priors outlined in [34], who considered regression settings only, carries over to density estimation; the techniques presented could also be used for other tree priors beyond OPTs.

The chapter is organized as follows. Section 3.2 introduces a class of tree–based priors on density functions, of which OPTs are a special case. Section 3.3 states our main result on tree–based supremum norm contraction, while Section 3.4 focuses on Uncertainty Quantification, both for the density function and smooth functionals thereof. Section 3.5 illustrates our findings numerically through a simulation study. Section 3.6 briefly summarises and discusses the results and future research directions. Proofs are gathered in Sections 3.7 and 3.8.

3.2 Dyadic tree–based random densities and Optional Pólya trees (OPTs)

3.2.1 Bayesian framework

Adopting a Bayesian point of view, the density estimation model on $[0, 1)$ consists in observing

$$\begin{aligned} X = (X_1, \dots, X_n) \,|\, f &\sim P_f^{\otimes n} \\ f &\sim \Pi, \end{aligned} \tag{3.1}$$

where P_f is the distribution on $[0, 1)$ with density f with respect to Lebesgue measure: $dP_f = f d\mu$, and where Π is a prior distribution on densities f to be defined below. The posterior distribution is then the conditional distribution of f given X and is denoted $\Pi[\cdot \,|\, X]$.

Frequentist analysis of Bayesian posteriors. To analyse mathematically the behaviour of the posterior distribution $\Pi[\cdot \,|\, X]$, once the posterior is formed using the Bayesian model, we make the frequentist assumption that the data X has actually been generated from a 'true' parameter value f_0, that is, in the density estimation setting, $X \sim P_{f_0}^{\otimes n}$. In the sequel, we thus study the behaviour of $\Pi[\cdot \,|\, X]$ in probability under $P_{f_0} = P_{f_0}^{\otimes n}$. For more details and context, we refer the reader to the book [65].

Motivated by recent work [34] on Bayesian CART in regression settings (see e.g. the discussion in Section 5 of [34]), we introduce a family of tree-based prior distributions on density functions. For simplicity, we mostly consider the case of densities on the unit interval, but our results could be extended to higher dimensions up to using slightly more complex notation, which we refrain to do here – see, though, the discussion in Section 3.6 for more on this –.

Informal prior description. The prior on densities is defined in three steps, which will be more formally introduced below

Step 1 a random tree $\mathcal{T}$ is sampled from a prior $\Pi_{\mathbb{T}}$ on trees;

Step 2 given $\mathcal{T}$, a partition $I_{\mathcal{T}}$ of the unit interval is produced, built recursively in a tree fashion 'along' $\mathcal{T}$ with breakpoints placed at midpoints of the successive intervals;

Step 3 given $I_{\mathcal{T}}$, the output density f is a histogram with random heights whose distribution follows a Pólya tree–type law.

3.2.2 Priors $\Pi_{\mathbb{T}}$ on full binary trees

Definition 3. *A* full binary tree *is a set of nodes* $\mathcal{T} = \left\{(l,k),\ l \geq 0,\ 0 \leq k \leq 2^l - 1\right\}$ *verifying the condition*

$$(l,k) \in \mathcal{T} \implies \text{if } l > 0,\ \left(l-1, \lfloor k/2 \rfloor\right) \in \mathcal{T} \text{ and } \left(l, k + (-1)^k\right) \in \mathcal{T}.$$

One then says that $\left(l-1, \lfloor k/2 \rfloor\right)$ *is the* parent *node of its* children (l,k) *and* $\left(l, k+(-1)^k\right)$*, and a node with no children is called an external node or leaf;* $(0,0)$ *belongs to every non-empty tree and is called the tree* root. *We denote by* $\mathcal{T}_{int}$ *the set of non-terminal – or 'internal' – nodes in* $\mathcal{T}$ *(i.e. those with children), and* $\mathcal{T}_{ext} = \mathcal{T} \setminus \mathcal{T}_{int}$ *the set of 'leaves' – also called 'external' nodes –.*

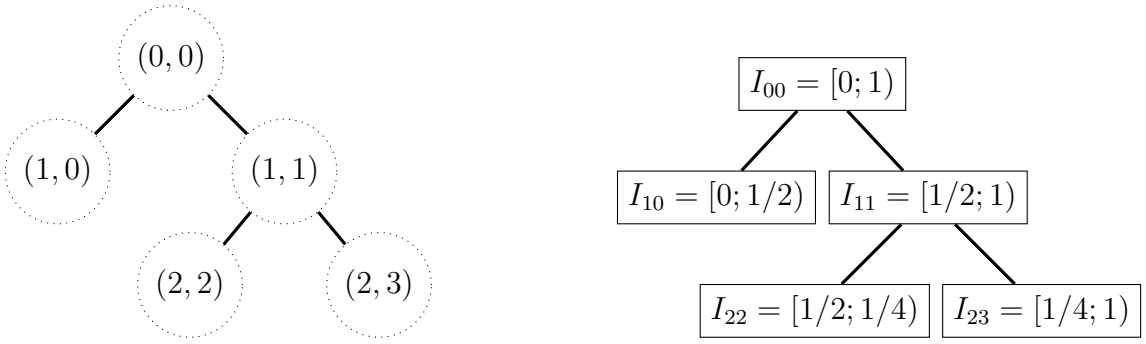
The parent-child relationship of the pairs in a tree gives rise to the tree representation depicted on Figure 3.1a. This justifies the following terminology as we define the *depth* of $\mathcal{T}$ as the integer

$$d(\mathcal{T}) \coloneqq \max_{(l,k)\in\mathcal{T}} l.$$

One further denotes by $\mathbb{T}$ the set of all binary trees and, putting a slight restriction on the maximum depth,

$$\mathbb{T}_n \coloneqq \{\mathcal{T} \in \mathbb{T} :\ d(\mathcal{T}) \leq L_{\max}\}, \qquad \text{with } L_{\max} \coloneqq \left\lfloor \log_2\left(n/\log^2(n)\right) \right\rfloor. \tag{3.2}$$

The prior distributions considered below put mass 1 to the subset $\mathbb{T}_n$ of $\mathbb{T}$.



(a) Tree pairs.

(b) Tree partitioning $I_{\mathcal{T}}$.

Figure 3.1: Tree $\mathcal{T} = \{(0,0),(1,0),(1,1),(2,2),(2,3)\}$.

Next we give two examples of priors $\Pi_{\mathbb{T}}$ on full binary trees. Both are actually considered in actual Bayesian CART implementations [40, 49].

Example 3 ($\mathrm{GW}(p)$ Markov process on tree)**.** *A random tree is recursively defined by the following process. First, let us attribute to each possible pair (l,k) a deterministic parameter $p_{lk} \in [0,1]$. Starting at the root node $(0,0)$, either the tree with only $(0,0)$ as node is returned with probability $1-p_{00}$, or there is a split and the tree contains not only $(0,0)$ but at least also $(1,0)$ and $(0,1)$. The construction process then continues recursively until either there are no further nodes to split, or a maximum depth L_{max} is reached, after which (i.e. for $l \geq L_{max}$) we do not further grow the tree. More precisely, the recursion is from up to down (l grows) and left to right (k grows), as follows: given the tree contains (l,k), with probability $1-p_{lk}$ the node (l,k) is a leaf; and with probability p_{lk}, the tree further has a split at (l,k), i.e. the node (l,k) has $(l+1,2k)$ and $(l+1,2k+1)$ as children in the tree.*

The process producing such a random tree $\mathcal{T}$ is Markov (along the complete dyadic tree) in the sense that the probability that a node (l,k) further splits only depends on the fact that the node is present or not and on the parameter p_{lk}, but not on the rest of the tree built so far (above and to the left of (l,k)). By analogy to Galton–Watson processes, with here nodes having either two or zero children with probabilities p_{lk} and $1-p_{lk}$ respectively, we call $\Pi_{\mathbb{T}}$ as above a $\mathrm{GW}(p)$ prior, with parameters $p=(p_{lk})=(p_\epsilon)$ (we define the link between ϵ and (l,k) below, in Section 3.2.3), $p_{L_{max}k}=0$.

Example 4 (Conditioning on the number of leaves)**.** *In this construction, one samples first a number K of leaves according to a prior on integers and given K one then samples uniformly from the set of all full binary trees with K leaves and depth at most L_{max}.*

3.2.3 Partitioning $I_{\mathcal{T}}$

Let us first introduce notation on dyadic numbers and intervals. For any binary sequence $\epsilon \in \{0,1\}^l$, its length is $|\epsilon| = l > 0$. For any dyadic number $r = k/2^l$ in $[0,1)$ with $0 \leq k < 2^l,\ l > 0$, one writes $\epsilon(k,l) = \epsilon_1(r)\cdots\epsilon_l(r) \in \{0,1\}^l$, such that $r = \sum_{k=1}^{l} \epsilon_k(r)2^{-k}$, its unique decomposition in base 2^{-1} with $|\epsilon| = l$. Accordingly, one introduces the dyadic intervals, for $\epsilon = \epsilon(k,l)$,

$$I_\epsilon := I_{lk} := \left[\frac{k}{2^l}, \frac{k+1}{2^l}\right),$$

and one sets $I_\varnothing = I_{0,0} = [0,1)$. In addition, for any ϵ and $0 < i \leq |\epsilon|$, one writes $\epsilon^{[i]} = \epsilon_1 \ldots \epsilon_i$. Also, we introduce $\mathcal{E}^* = \cup_{l=0}^{\infty} \{0;1\}^l$ where $\{0;1\}^0 = \{\varnothing\}$.

To each full binary tree encoded as above as the collection of its nodes (l,k), we associate a partition $I_{\mathcal{T}}$ of the unit interval given by, with $\mathcal{T}_{ext}$ the external nodes of $\mathcal{T}$ as in Definition 3,

$$[0,1) = \bigcup_{(l,k)\in\mathcal{T}_{ext}} I_{lk}.$$

Such a tree-based recursive partitioning of $[0,1)$ is illustrated on Figure 3.1b. The deeper the tree locally, the more refined the corresponding partition becomes. By definition of I_{lk}, note that the partition has split-points at dyadic numbers. The final partition $I_{\mathcal{T}}$ can also be seen as being obtained from recursively splitting $[0,1)$ in halves, continuing to split locally only if the tree continues further down at that location. For this reason we talk about *splitting at midpoints*. Note that, still using full binary trees $\mathcal{T}$, one could make splits at a different, possibly random, location. Although this makes the construction even more flexible, we shall not consider this here for simplicity (we note in passing that computationally the

split–at–midpoint construction appears often to be among the easiest to simulate from, as it does not require to draw split locations; we refer to [34], Section 4, for more on 'unbalanced' splits).

3.2.4 Prior values given tree and partitioning

Once a tree $\mathcal{T}$ and partitioning $I_{\mathcal{T}}$ are given, we draw a random histogram over the partition given by $I_{\mathcal{T}}$ by sampling heights over each sub-interval in such a way that the overall histogram is a positive density f (i.e. $f > 0$ and $\int_0^1 f = 1$). To do so, we use a mass–splitting process along the tree $\mathcal{T}$, which actually coincides with that of Pólya trees – we refer to the Appendix 3.8.1 for more on those –. This choice is for simplicity but we could consider other choices too (in this vein, the $\mathrm{Beta}(a, a)$ law at the end of Definition 4 could be taken to depend on (l, k) or be a different distribution).

Definition 4 (Prior Π)**.** *Let $\Pi_{\mathbb{T}}$ be a prior on full binary trees. Let (Y_ε) be a sequence of independent variables of distribution $\mathrm{Beta}(a_{\varepsilon 0}, a_{\varepsilon 1})$, for some $a_{\varepsilon 0}, a_{\varepsilon 1} \in [0, 1]$, indexed by $\varepsilon \in \mathcal{E}^*$. The prior Π draws a random tree–based histogram f as follows*

$$\mathcal{T} \sim \Pi_{\mathbb{T}} \tag{3.3}$$

$$f \,|\, \mathcal{T} \sim \sum_{\varepsilon \equiv (l,k) \in \mathcal{T}_{ext}} h_\varepsilon \mathbb{1}_{I_{lk}}, \qquad \text{with} \quad h_\varepsilon = 2^l \prod_{i=1}^{l} Y_{\varepsilon[i]}. \tag{3.4}$$

The distribution $f \,|\, \mathcal{T} = T$ for a given $T \in \mathbb{T}$ is called a T–Pólya tree with parameters (a_ε). In the sequel we set $a_\varepsilon = a$ for some fixed $a > 0$, in which case the distribution is denoted as T–$\mathrm{PT}(a)$.

It results from the definition that the overall prior Π is a mixture of T–Pólya trees. When the mixing distribution $\Pi_{\mathbb{T}}$ is a $\mathrm{GW}(p)$ prior, it turns out that Π coincides with Optional Pólya trees introduced in [173], in the case of splits at midpoints.

Proposition 1. *Let Π be the mixture distribution induced on densities f constructed as*

$$\mathcal{T} \sim \mathrm{GW}(p)$$
$$f \,|\, \mathcal{T} \sim \mathcal{T}\text{–}\mathrm{PT}(a).$$

Then Π coincides with the Optional Pólya tree of [173] corresponding to the recursive partitioning $\{I_\epsilon,\ \epsilon \in \mathcal{E}^*\}$ with splits at midpoints and parameters $M(I_\epsilon) = \lambda(I_\epsilon) = 1, K_1(I_\epsilon) = 2$, stopping probabilities $\rho(I_\epsilon) = 1 - p_\epsilon$ for any $\epsilon \in \mathcal{E}^*$ and parameters for mass allocation $\alpha_1^1 = \alpha_1^2 = a$.

The proof of Proposition 1 is presented in Appendix 3.8.2. Our notation differs slightly from [173] (which does not make the tree connection) for two reasons: first, the tree–setting enables one to use the framework of [34] and second, although in what follows we stick to OPTs for simplicity, the same proofs work nearly unmodified for other tree–priors, such as the one in Example 4.

3.2.5 Posterior distribution

Let us recall that the prior Π in Definition 4 is the mixture

$$\begin{aligned} \mathcal{T} &\sim \Pi_{\mathbb{T}} \\ f \,|\, \mathcal{T} &\sim \Pi(\cdot \,|\, \mathcal{T}), \end{aligned} \tag{3.5}$$

where $\Pi(\cdot \,|\, \mathcal{T})$ is, given $\mathcal{T}$, a $\mathcal{T}$–Pólya tree. For a given dyadic interval I, let $N_X(I)$ denote the number of points X_i that fall in I. The next result is proved in Appendix 3.8.3.

Proposition 2 (Posterior given $\mathcal{T}$)**.** *Suppose the prior is given by* (3.5)*, where the prior given $\mathcal{T}$ is a $\mathcal{T}$–Pólya tree with parameters (a_ε). Then, in the density estimation model* (3.1)*, the posterior $\Pi[\cdot \,|\, X, \mathcal{T}]$ is a $\mathcal{T}$–Pólya tree with parameters (a_ε^X) given by, for any $\varepsilon \in \mathcal{E}^*$,*

$$a_\varepsilon^X = a_\varepsilon + N_X(I_\varepsilon).$$

Let us now move on to describe the posterior induced on trees. We denote

$$N_T(X) = \int \prod_{i=1}^n f(X_i) d\Pi(f \,|\, \mathcal{T} = T) \tag{3.6}$$

the marginal distribution of X given $\mathcal{T} = T$. It follows from Bayes' formula that $\Pi[\cdot \,|\, X]$ induces a posterior distribution on trees given as: for any $T \in \mathbb{T}$, and $N_T(X)$ as in (3.6),

$$\Pi[\mathcal{T} = T \,|\, X] = \frac{\Pi_{\mathbb{T}}[\mathcal{T} = T] N_T(X)}{\sum_{T \in \mathbb{T}} \Pi_{\mathbb{T}}[\mathcal{T} = T] N_T(X)}. \tag{3.7}$$

This is in general a fairly complicated distribution with no closed–form expression. In case the prior $\Pi_{\mathbb{T}}$ on trees is $\mathrm{GW}(p)$, it turns out that the posterior on trees is $\mathrm{GW}(p^X)$ for updated parameters p^X. Let, for $a > 0$,

$$\nu_\varepsilon^X = 2^{N_X(I_\varepsilon)} \frac{B(a + N_X(I_{\varepsilon 0}), a + N_X(I_{\varepsilon 1}))}{B(a, a)}. \tag{3.8}$$

Let us now consider parameters (p_ε^X) given by the equations

$$\frac{p_\varepsilon^X}{1 - p_\varepsilon^X}(1 - p_{\varepsilon 0}^X)(1 - p_{\varepsilon 1}^X) = \frac{p_\varepsilon}{1 - p_\varepsilon}(1 - p_{\varepsilon 0})(1 - p_{\varepsilon 1})\nu_\varepsilon^X, \tag{3.9}$$

Equations (3.9) together admit a unique solution (p_ε^X) obtained by a bottom–up recursion noting that for $|\varepsilon| = L_{\mathsf{max}}$, $p_\varepsilon^X = p_\varepsilon = 0$. This is verified along the proof of Proposition 3 below.

Proposition 3 (Special case of OPTs)**.** *In the setting of Proposition 2, suppose further that the distribution $\Pi_{\mathbb{T}}$ on trees is $\mathrm{GW}(p)$ with split probabilities (p_ε). Then the posterior distribution can be described as*

$$\begin{aligned} \Pi[\mathcal{T} = \cdot \,|\, X] &\sim \mathrm{GW}(p_\varepsilon^X) \\ \Pi[\cdot \,|\, X, \mathcal{T}] &\sim \mathcal{T}\text{–PT}(a_\varepsilon^X) \end{aligned}$$

with splits probabilities (p_ε^X) verifying the recursion (3.9) *and a_ε^X as in Proposition 2. In other words the posterior follows an OPT distribution with corresponding hyperparameters as specified in Proposition 1.*

The proof of this proposition is presented in Appendix 3.8.3.

3.2.6 Notation and function spaces

Below we shall consider the Hölder class of functions with support in $[0,1)$ and smoothness parameter $0 < \alpha \leq 1$, defined as

$$\mathcal{C}^{\alpha}[0,1) \coloneqq \left\{ f : [0,1) \mapsto \mathbb{R}, \quad \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x-y|^{\alpha}} < +\infty \right\}$$

and we similarly define Hölder balls with parameters $\alpha > 0$ and $K \geq 0$ as

$$\Sigma(\alpha, K) \coloneqq \left\{ f : [0,1) \mapsto \mathbb{R}, \quad \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x-y|^{\alpha}} \leq K \right\}.$$

Bounded Lipschitz metric. Let $(\mathcal{S}, d)$ be a metric space. The bounded Lipschitz metric $\beta_{\mathcal{S}}$ on probability measures of $\mathcal{S}$ is defined as, for any μ, ν probability measures of $\mathcal{S}$,

$$\beta_{\mathcal{S}}(\mu, \nu) = \sup_{F; \|F\|_{BL} \leq 1} \left| \int_{\mathcal{S}} F(x)(d\mu(x) - d\nu(x)) \right|, \tag{3.10}$$

where $F : \mathcal{S} \to \mathbb{R}$ and

$$\|F\|_{BL} = \sup_{x \in \mathcal{S}} |F(x)| + \sup_{x \neq y} \frac{|F(x) - F(y)|}{d(x,y)}. \tag{3.11}$$

This metric metrises the convergence in distribution, see e.g. [55], Theorem 11.3.3.

As shown in [29], it is also useful to introduce the Haar wavelet basis to carry out an analysis of Pólya tree-like posterior distributions. Indeed, one can relate the inclusion of a node (l,k) in a tree $\mathcal{T}$ to the fact that the coefficient corresponding to the Haar wavelet function ψ_{lk} in the decomposition of $f \sim \Pi[\cdot|\mathcal{T}]$ is non-zero almost surely. More precisely, the Haar basis of $L^2[0;1)$ is the family composed of the mother wavelet $\phi = \mathbb{1}_{[0;1)}$ and the functions

$$\psi_{lk}(\cdot) = 2^{l/2} \psi(2^l \cdot - k)$$

for $l \geq 0$ and $0 \leq k < 2^l$, where $\psi = \mathbb{1}_{[1/2;1)} - \mathbb{1}_{[0;1/2)}$. However, as we consider the problem of density estimation, maps f under scrutiny all verify $\langle f, \phi \rangle = \int_0^1 f(t)dt = 1$, so that we only focus on the wavelets ψ_{lk} and the corresponding coefficients $f_{lk} \coloneqq \langle f, \psi_{lk} \rangle$ in the following. As for the true density, we define $f_{0,lk} \coloneqq \langle f_0, \psi_{l,k} \rangle$.

3.3 Posterior contraction rates for OPTs

For any $\alpha > 0$, $\mu > 0$, $K \geq 0$, we define the regularity class of densities

$$\mathcal{F}(\alpha, K, \mu) \coloneqq \left\{ f \geq \mu, \quad \int_0^1 f = 1, \quad f \in \Sigma(\alpha, K) \right\},$$

as well as the sequence

$$\varepsilon_n(\alpha) \coloneqq \left(n^{-1} \log^2 n \right)^{\frac{\alpha}{1+2\alpha}}. \tag{3.12}$$

Up to a logarithmic factor, this corresponds to the minimax supremum norm rate of estimation over the class $\mathcal{F}(\alpha, K, \mu)$, which equals $(n/\log n)^{-\alpha/(1+2\alpha)}$ up to constants [80].

3.3.1 Supremum norm convergence for the whole posterior distribution

We now show that the posterior distribution $\Pi[\cdot \,|\, X]$ asymptotically concentrates most of its mass on a $\|\cdot\|_\infty$–ball of optimal radius.

Theorem 11. *Suppose that $f_0 \in \mathcal{F}(\alpha, K, \mu)$ for some $\mu > 0$, $0 < \alpha \leq 1$ and $K \geq 0$. Let Π be an OPT prior with split probabilities $p_{lk} = \Gamma^{-l}$, $l \geq 0$, $0 \leq k < 2^l$, $\Gamma > 0$, and parameter $a > 0$. Then, for Γ large enough, any sequence $M_n \to \infty$, as $n \to \infty$, and $\varepsilon_n = \varepsilon_n(\alpha)$ as in (3.12),*

$$E_{f_0}\Pi\Big[\, \|f - f_0\|_\infty > M_n \varepsilon_n \,|\, X\Big] \to 0.$$

Theorem 11 shows that an OPT posterior with split probabilities decreasing exponentially fast with nodes depth concentrates most of its mass in a supremum norm ball of (near–) minimax optimal radius, whenever the signal has regularity $\alpha \leq 1$. Some comments are in order. First, the regularity requirement $\alpha \leq 1$ is typical and expected for 'hard trees', which produce histogram-type estimators. An alternative would be to use 'soft trees', where individual learner are smooth [107, 34], see also the discussion in Section 3.6. Second, the slight loss of a logarithmic term in the convergence rate can be shown to be intrinsic to trees and is not due to a possible suboptimality of our rate upper–bounds: this has been formally shown in [34], Theorem 2, in a regression context; an analogous result could be shown in density estimation in a similar way.

A consequence of Theorem 11 is that a posterior draw is close with high probability to the true unknown density function of interest. This settles the *estimation* problem, but it does not yet say much about the quantification of uncertainty, i.e. the construction of *confidence sets*, a question addressed in Section 3.4.

3.3.2 Convergence rate for the median tree

While Theorem 11 entails convergence in probability of a draw from $\Pi[\cdot \,|\, X]$, one may ask what happens for aspects of such distribution, e.g. point estimators derived from it. A natural such estimator from the point of view of tree priors is the median tree estimator defined below, since there is a natural tree associated to it. Such an estimator will also turn helpful for uncertainty quantification as considered below.

The *median tree* is defined as the tree $\mathcal{T}^*$ whose interior nodes are

$$\mathcal{T}^*_{\text{int}} = \{(l,k) : \ \Pi[(l,k) \in \mathcal{T}_{\text{int}} | X] > 1/2\}\,, \tag{3.13}$$

and which is actually a tree as defined previously (see [34], Lemma 13). One associates to it the *median tree density estimator*

$$\hat{f}_{\mathcal{T}^*} = 1 + \sum_{(l,k)\in\mathcal{T}^*_{\text{int}}} 2^{l/2} \frac{N_X\left(I_{(l+1)(2k+1)}\right) - N_X\left(I_{(l+1)(2k)}\right)}{n} \psi_{lk}. \tag{3.14}$$

Lemma 27 in the appendix shows that this estimator converges in probability to the actual density f_0 at the same almost-minimax rate ε_n in supnorm as in Theorem 11. In Section 3.5, examples of $\mathcal{T}^*$ and $\hat{f}_{\mathcal{T}^*}$ are presented in Figures 3.2 and 3.3.

3.4 Uncertainty quantification for OPTs

In nonparametrics the problem of uncertainty quantification is well–known to be more delicate than the one of estimation: first negative results to the ambitious goal of constructing confidence sets that both cover the unknown truth and have a diameter that adapts in an optimal way to the smoothness of the unknown function or density were due to [105] and [112]. The general picture that emerged in recent years following these early works is that the difficulty of the problem depends on the considered loss function and on certain testing rates of separation, see [69], Chapter 8. Notably, for the supremum norm, contrary to L^2–losses for which some 'window' of adaptation is possible, constructing adaptive confidence sets in full generality is impossible unless one restricts the set of possible functions by assuming e.g. self–similarity conditions. Such conditions can be shown to be essentially necessary; they are also fairly natural from the practical perspective given that self–similarity is itself quite wide–spread in natural phenomena.

Let us briefly describe the uncertainty quantification results we derive. A first confidence band based on the posterior median and using self–similarity is built in Section 3.4.2. Next, we prove in Section 3.4.3 that the quantile posterior credible set for the cumulative distribution function leads to optimal UQ; this is a consequence of a more general result, an (adaptive) nonparametric Bernstein–von Mises theorem, proved in Appendix 3.8.5. Finally in Section 3.4.4 we construct a confidence band integrating further information from some functionals that is less conservative than the simple band constructed in Section 3.4.2 and achieves a target confidence level. Our results can be seen as counterparts in density estimation and for tree priors of the results in [139]. Another approach in density estimation would be to use spike–and–slab Pólya priors as recently considered by the second author in [31]. Nevertheless, the latter are expected to be less efficient to compute in high–dimensions (as they, e.g., require to explore all wavelet coefficients in the different dimensions), a setting that, while not investigated in the present chapter, is particularly promising for OPTs, see also the discussion in Section 3.6.

3.4.1 A self-similarity condition

Here we take the same condition as in [139] (see also [69]). It is fairly simple to state, and can be only slightly improved (see [22]).

Definition 5 (Set $\mathcal{S}$ of self–similar functions)**.** *Given an integer $j_0 > 0$ and $\alpha \in (0, 1]$, we say that $f \in \Sigma(\alpha, K)$ is* self-similar *if, for some constant $\eta > 0$,*

$$\|K_j(f) - f\|_\infty \geq \eta 2^{-j\alpha} \text{ for all } j \geq j_0,$$

where $K_j(f) = \sum_{l<j} \sum_k \langle f, \psi_{lk}\rangle \psi_{lk}$. The set of such f's is denoted $\mathcal{S} = \mathcal{S}(\alpha, K, \eta)$.

The condition assumes that at each resolution depth $j \geq j_0$, the overall 'energy' (measured in terms of supremum norm) of the wavelet coefficients at levels larger than j is lower bounded by a typical amount for α–Hölder functions. Indeed, for any $j \geq j_0$, the quantity $\|K_j(f) - f\|_\infty$ is itself also upper–bounded up to a constant by the same quantity (this follows from standard bounds on the supremum norm and the definition of the Hölder class).

3.4.2 Simple confidence band

A first construction consists in defining a band from a centering function and a radius. A first and simple possibility consists in defining those using the median tree (3.13): the resulting

median tree estimator (3.14) can serve as center, while a radius can be defined as

$$\sigma_n = v_n \sqrt{\frac{\log n}{n}} 2^{d(\mathcal{T}^*)/2}, \tag{3.15}$$

where $d(\mathcal{T}^*)$ is the depth of the median tree $\mathcal{T}^*$, for some slowly diverging sequence (v_n) as specified below. This allows us to define the confidence band, for $\hat{f}_{\mathcal{T}^*}$ as in (3.14),

$$\mathcal{C}_n = \left\{ f : \left\| f - \hat{f}_{\mathcal{T}^*} \right\|_\infty \leq \sigma_n \right\}. \tag{3.16}$$

Under self–similarity as in Definition 5, the median tree can in particular be shown to have a depth of the order of the oracle cut–off $2^{L_n^*} \approx n^{1/(2\alpha+1)}$ (up to a logarithmic factor, see the Appendix for a precise statement in Lemma 25) which in turn implies desirable properties for the band $\mathcal{C}_n$ as is made explicit in the next theorem.

Theorem 12. *Let $0 < \alpha_1 < \alpha_2 \leq 1$, $K > 0$, $\mu > 0$ and $\eta > 0$. Let Π be the same prior as in Theorem 11, $\mathcal{C}_n$ as in (3.16) with $v_n / \log^{1/2} n \to \infty$, then uniformly on $f_0 \in \mathcal{S}(\alpha, K, \eta) \cap \mathcal{F}(\alpha, K, \mu)$, $\alpha \in [\alpha_1, \alpha_2]$,*

$$|\mathcal{C}_n|_\infty = O_{P_0}\left(v_n \left(\frac{\log n}{n} \right)^{\alpha/(2\alpha+1)} \right)$$

and

$$P_0 \left[f_0 \in \mathcal{C}_n \right] = 1 + o(1), \qquad \Pi[\mathcal{C}_n \,|\, X] = 1 + o_{P_0}(1).$$

For a slowly diverging sequence (v_n), the diameter of $\mathcal{C}_n$ is then within a logarithmic factor of the minimax rate of estimation on $\Sigma(\alpha, K)$ with high probability. It is attained adaptively (the definition of $\mathcal{C}_n$ does not depend on α) for any window $[\alpha_1; \alpha_2]$. The set $\mathcal{C}_n$ allows to quantify uncertainty on f_0 as it is an asymptotic confidence set, and it is also a credible set of credibility going to 1.

3.4.3 UQ for functionals: a Donsker–type theorem

OPTs with flat initialisation. Let us introduce a slight modification of the OPT prior where trees from the prior distribution are constrained to include all nodes of depth less than some number $l_0 = l_0(n)$, slowly diverging to ∞.

Definition 6. *A prior on densities Π of the type (3.5) is said to have* flat initialisation up to level $l_0 = l_0(n)$ *if the prior on trees $\Pi_{\mathbb{T}}$ verifies*

$$\Pi_{\mathbb{T}} \left[\bigcap_{l \leq l_0(n), k} \{(l,k) \in \mathcal{T}\} \right] = 1.$$

The next result considers the behaviour of the induced posterior on $F(\cdot) = \int_0^\cdot f$, that is on the distribution function for an OPT prior on f. Let us also define, for $\hat{f}_{\mathcal{T}^*}$ the median tree estimator,

$$\hat{F}_n^{med}(t) = \int_0^t \hat{f}_{\mathcal{T}^*}(u) du. \tag{3.17}$$

Let us recall that for Q a probability measure on $[0,1]$ of distribution function H, a Q–Brownian bridge is a centered Gaussian process $Z(t)$ with covariance function $E[Z(s)Z(t)] = \min(H(s), H(t)) - H(s)H(t)$ and $0 \le s, t \le 1$.

Theorem 13 (Donsker's theorem for OPTs). *Let $X = (X_1, \dots, X_n)$ be i.i.d. from law P_0 with density f_0. Let $f_0 \in \mathcal{F}(\alpha, K, \mu)$, for some $\alpha \in (0;1]$, $K \ge 0$, $\mu > 0$. Let Π be an OPT prior with flat initialisation up to level $l_0(n)$ that verifies $\sqrt{\log n} \le l_0(n) \le \log n / \log\log n$, and other than that for $l > l_0(n)$ with same parameters as the prior in Theorem 11.*

Let G_{P_0} be a P_0-Brownian bridge $G_{P_0}(t), t \in [0,1)$. For $\hat{F}_n^{med}$ as in (3.17), as $n \to \infty$,

$$\beta_{C[0,1)}\left(\mathcal{L}(\sqrt{n}(F - \hat{F}_n^{med}) \,|\, X), \mathcal{L}(G_{P_0})\right) \to^{P_{f_0}} 0.$$

Furthermore, for F_n the empirical distribution function, as $n \to \infty$,

$$\beta_{L^\infty[0,1)}\left(\mathcal{L}(\sqrt{n}(F - F_n) \,|\, X), \mathcal{L}(G_{P_0})\right) \to^{P_{f_0}} 0.$$

This implies that the induced posterior distribution $\mathcal{L}(\sqrt{n}\|F - \hat{F}_n^{med}\|_\infty \,|\, X)$ converges weakly in probability to $\mathcal{L}(\|G_{P_0}\|_\infty)$. Furthermore, for $0 < \gamma < 1$, the credible set

$$\mathcal{F}_n = \{F : \ \|F - \hat{F}_n^{med}\|_\infty \le \rho_n^X\},$$

with ρ_n^X chosen such that $\Pi[\mathcal{F}_n \,|\, X] = 1 - \gamma$, is an asymptotically optimal (efficient) confidence set of level $1 - \gamma$. We refer to [33] for more details on this; note that in the latter paper the results are for priors of fixed regularity only, whereas here the prior additionally enables adaptation to the smoothness of f. The behaviour of the credible set $\mathcal{F}_n$ is illustrated in Figure 3.5.

3.4.4 Multiscale confidence band

Here we follow the approach introduced in [32, 33] and first briefly recall the idea. One wishes to define a 'multiscale' space (i.e. defined from wavelet coefficients) with an associated metric that is weak enough so that convergence of the posterior distribution for f in that space converges at rate $1/\sqrt{n}$, instead of the slower nonparametric rate of order $n^{-\alpha/(2\alpha+1)}$. In such space one can then formulate a convergence of the posterior to a Gaussian limit, namely a nonparametric Bernstein–von Mises theorem. Below we only define the multiscale space as it is used in the definition of the credible band and postpone details on the precise statement of convergence to Appendix 3.8.5.

Let us call the sequence $w = (w_l)_{l \ge 0}$ 'admissible' if $w_l / \sqrt{l} \to \infty$ as $l \to \infty$. For such a sequence, let us define

$$\mathcal{M}_0 = \mathcal{M}_0(w) = \left\{ x = (x_{lk})_{l,k}, \ \lim_{l \to \infty} \max_{0 \le k < 2^l} \frac{|x_{lk}|}{w_l} = 0 \right\}. \tag{3.18}$$

Equipped with the norm $\|x\|_{\mathcal{M}_0} = \sup_{l \ge 0} \max_{0 \le k < 2^l} |x_{lk}| / w_l$, this is a separable Banach space [33]. In a slight abuse of notation, we write $f \in \mathcal{M}_0$ if the sequence of its Haar wavelet coefficients $(\langle f, \psi_{lk} \rangle)_{l,k}$ belongs to that space.

Let us consider a credible ball in the space $\mathcal{M}_0$: recalling the definition (3.14) of the median tree estimator $\hat{f}_{\mathcal{T}^*}$, let us choose $R_n = R_n(X)$ in such a way that

$$\Pi[\|f - \hat{f}_{\mathcal{T}^*}\|_{\mathcal{M}_0(w)} \le R_n / \sqrt{n} \,|\, X] = 1 - \gamma \tag{3.19}$$

(or possibly $\geq 1-\gamma$ if the equation has no exact solution, in which case the limit in the confidence statement of the next proposition is replaced by a liminf and equality by $\geq$).

Let us define, for R_n as in (3.19), σ_n as in (3.15) and $f_{\mathcal{T}^*}$ the median tree estimator (3.14),

$$\mathcal{C}_n^{\mathcal{M}} = \left\{f : \ \|f - \hat{f}_{\mathcal{T}^*}\|_\infty \leq \sigma_n\right\} \bigcap \left\{f : \ \|f - \hat{f}_{\mathcal{T}^*}\|_{\mathcal{M}_0(w)} \leq R_n/\sqrt{n}\right\}. \tag{3.20}$$

The next result states that $\mathcal{C}_n^{\mathcal{M}}$ is under self–similarity asympotically a confidence band of prescribed level $1-\gamma$.

Proposition 4. *Let $0 < \alpha_1 < \alpha_2 \leq 1$, $K > 0$, $\mu > 0$ and $\eta > 0$. Let $\mathcal{C}_n^{\mathcal{M}}$ be defined by (3.20), for $v_n/\log^{1/2} n \to \infty$, and Π an OPT prior with flat initialisation up to level $l_0(n)$ that verifies $\sqrt{\log n} \leq l_0(n) \leq \log n/\log\log n$, and other than that for $l > l_0(n)$ with same parameters as the prior in Theorem 11. First, the set $\mathcal{C}_n^{\mathcal{M}}$ is a $(1-\gamma)-$credible band as, uniformly on $\alpha \in [\alpha_1, \alpha_2]$ and $f_0 \in \mathcal{S}(\alpha, K, \eta) \cap \mathcal{F}(\alpha, K, \mu)$,*

$$\Pi[\mathcal{C}_n^{\mathcal{M}} \,|\, X] = 1 - \gamma + o_{P_0}(1).$$

Further, under the same conditions,

$$\left|\mathcal{C}_n^{\mathcal{M}}\right|_\infty = O_{P_0}\left(v_n\left(\frac{\log n}{n}\right)^{\alpha/(2\alpha+1)}\right),$$
$$P_0\left[f_0 \in \mathcal{C}_n^{\mathcal{M}}\right] = 1 - \gamma + o(1).$$

Proposition 4 quite directly follows from combining Theorem 12, which concerns $\mathcal{C}_n$ and the nonparametric BvM Theorem 14 proved in the Appendix, which concerns the second part of the intersection in (3.20). Compared to $\mathcal{C}_n$ the advantage of $\mathcal{C}_n^{\mathcal{M}}$ is that it uses more 'posterior information' by intersecting with the $\mathcal{M}_0(w)$ credible ball, resulting in a credible ball with both credibility and confidence close to a given user–specified confidence level $1-\gamma$. By contrast, $\mathcal{C}_n$ was more 'conservative' in this respect, having credibility and confidence both going to 1. The behaviour of the credible band $\mathcal{C}_n^{\mathcal{M}}$, in particular in comparison to $\mathcal{C}_n$ from (3.16), is illustrated in simulations in the following Section 3.5.

3.5 Simulation study

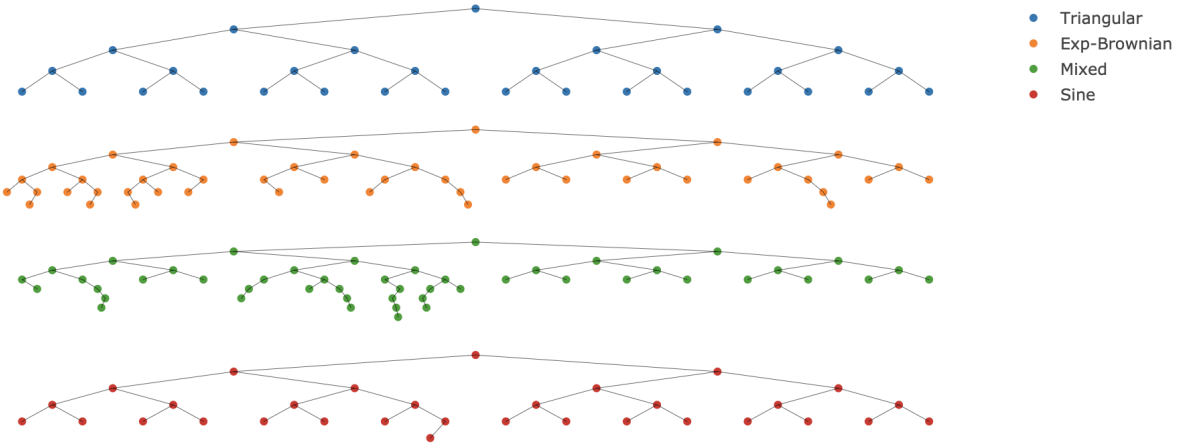
We consider the credible sets $\mathcal{C}_n$ and $\mathcal{C}_n^{\mathcal{M}}$ defined in (3.16) and (3.20) respectively and illustrate their coverage and diameter properties numerically through a simulated study.

We focus on a prior as in Proposition 4, with parameters $\Gamma = 1.1$, $a = 1$ and $l_0(n) = \sqrt{\log n}$. We take four fairly different densities f_0, illustrating different aspects of inference and UQ with Optional Pólya trees:

- The triangular density $x \mapsto (.5 + 2 * x)1_{0 \leq x < 0.5} + (1.5 - 2 * (x - .5))1_{0.5 \leq x < 1}$ that is Lipschitz regular.
- The density
$$t \mapsto \frac{e^{W_t}}{\int_0^1 e^{W_s} ds}$$
where $(W_t)_{t \in [0;1)}$ is a Brownian motion that is almost surely $(1/2-\delta)$–Hölder regular for any $0 < \delta < 1/2$.

- The density
$$t \mapsto C\left(e^{W_t} 1_{0\leq t<0.5} + c 1_{0.5\leq x<1}\right\}$$
for $(W_t)_{t\in[0;1)}$ a Brownian motion and C, c real numbers such that this actually defines a continuous density function. In this case, the regularity is different and of a higher order on the second half of the interval.
- The sine density $t \mapsto 1 + 0.5 * \sin(2\pi x) \in C^\infty([0;1))$.

Figure 3.2: Interior nodes $\mathcal{T}^*_{\text{int}}$ of the median tree - $n = 10^5$.

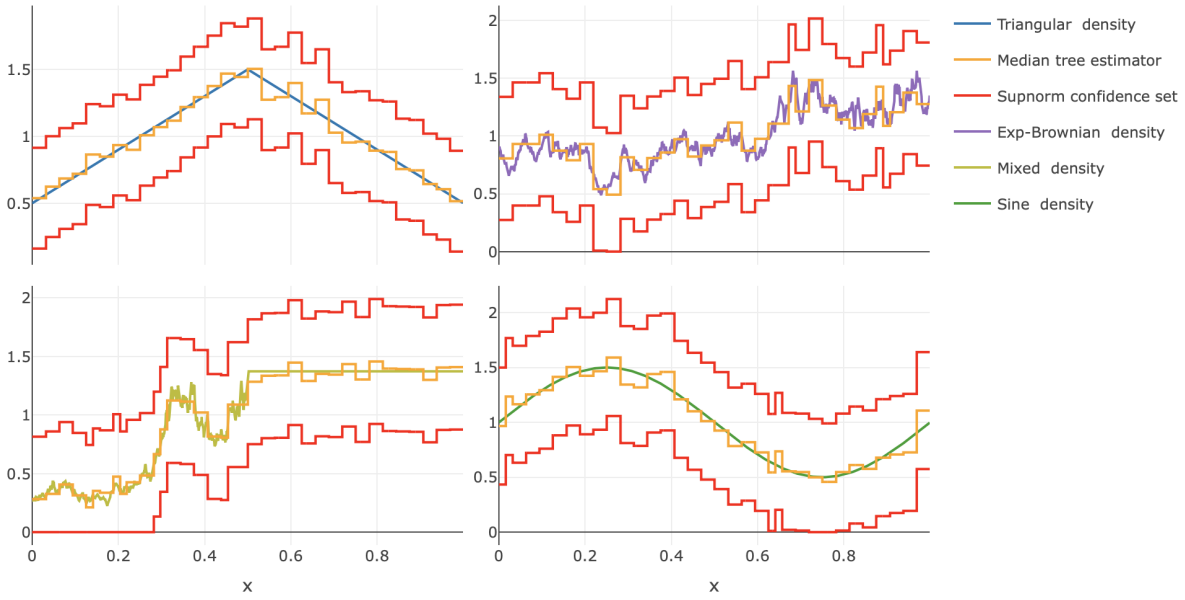


We first illustrate the behaviour of the median tree $\mathcal{T}^*$ and the associated estimator $\hat{f}_{\mathcal{T}^*}$ defined in (3.14) in these different situations. In Figure 3.2, we observe how this tree adapts to the regularity of the underlying sampling density f_0 via the interior nodes it selects. First, in the case of the smoother sine and triangular densities, fewer nodes are included, while the tree grows deeper with the other two more irregular signals. Indeed, as mentioned before and explicited in Lemma 25, the median tree can be shown to have a depth close to the oracle cut-off L_n^*, satisfying $2^{L_n^*} \approx n^{1/(2\alpha+1)}$. However, although the sine density is even more regular than the triangular one, their respective median trees have a similar behaviour and grow at the same pace. Indeed, since we use a piecewise constant tree estimator which relates to the Haar wavelet basis, our method cannot leverage additional regularities, beyond $\mathcal{C}^1[0,1)$. Finally, when it comes to the mixed density, the median tree has a spatial-dependent behaviour. It includes much more nodes in regions that corresponds to the first half of the sampling space, where the target regularity is that of the exp-Brownian density. As for the other half of the sampling space, it doesn't get deeper than $l_0(n)$. It highlights a desirable feature of tree-based methods, that is their spatial adaptivity. While we consider adaptation to global regularity in our theoretical results, one could also consider local adaptation, as was recently considered in [142], where results on local adaptation for tree–based priors (among others) are obtained in a regression setting.

In Figure 3.3, for the four sampling densities, we illustrate the estimator $\hat{f}_{\mathcal{T}^*}$ (orange) and the bounds of the credible set $\mathcal{C}_n$ (red), where we took $v_n = (\log n)^{0.501}$ in (3.15). The estimator (3.14) struggles to approximate the 'spiky' portions of the most irregular signals. Still, in any case, the credible band covers the true density f_0 as expected.

Then, to illustrate the intersected set $\mathcal{C}_n^{\mathcal{M}}$, defined in (3.20) via a multiscale condition, we sampled 10000 draws from the posterior and plotted, in Figure 3.4, 100 of those belonging to

Figure 3.3: Median tree estimator $\hat{f}_{\mathcal{T}^*}$ and credible set $\mathcal{C}_n$ - $n = 10^4$



Chosen significance γ	0.99	0.95	0.9	0.85
	$n = 10^4$			
Credibility of $C_n^{L^\infty}$	0.99	0.95	0.9	0.85
Credibility of $\mathcal{C}_n^{\mathcal{M}}$	0.99	0.95	0.8981	0.85
Credibility of $C_n^{L^\infty} \cap \mathcal{C}_n^{\mathcal{M}}$	0.9801	0.9029	0.8108	0.725
Credibility of the intersection if independence	0.9801	0.9025	0.81	0.7225
	$n = 10^5$			
Credibility of $C_n^{L^\infty}$	0.99	0.95	0.9	0.85
Credibility of $\mathcal{C}_n^{\mathcal{M}}$	0.9894	0.9494	0.8994	0.8494
Credibility of $C_n^{L^\infty} \cap \mathcal{C}_n^{\mathcal{M}}$	0.9801	0.9028	0.8118	0.7254
Credibility of the intersection if independence	0.9795	0.9019	0.8095	0.722

Table 3.1: Credibility of sets $C_n^{L^\infty}$ and $\mathcal{C}_n^{\mathcal{M}}$ for the triangular density f_0.

the confidence band (blue), for $\gamma = 0.05$. Most of those samples do not seem to lie close to the bounds of $\mathcal{C}_n$ which is consistent with the fact that $\mathcal{C}_n$, resp. $\mathcal{C}_n^{\mathcal{M}}$, has a posterior mass close to 1, respectively 0.95. Though our illustrations concern the intersection of $\mathcal{C}_n^{\mathcal{M}}$ with the support the posterior, via the representation of posterior draws, it appears that $\mathcal{C}_n^{\mathcal{M}}$ is actually smaller than $\mathcal{C}_n$.

As for the confidence sets $\mathcal{F}_n$ on the cumulative distribution function $F_0(\cdot) = \int_0^{\cdot} f_0(t)dt$, we illustrate an example in Figure 3.5 for a smaller sample size of $n = 10^3$ and $\gamma = 0.95$. The bounds of $\mathcal{F}_n$ follow tightly the true signal and the set covers it, in spite of the fewer number of observations available compared to previous plots. Indeed, following the discussion after Theorem 13, $\mathcal{F}_n$ has a radius decreasing at the parametric rate $\sqrt{n}^{-1}$.

We end this section with an illustration of a phenomenon that was noticed and established in [139] for a spike-and-slab prior in a regression setting. Namely, since we constructed an adaptive $(1-\gamma)$-confidence bands whose diameter in supnorm shrinks at an almost optimal rate, one may wonder how much it differs from the $(1-\gamma)$-credible band in the supremum norm $C_n^{L^\infty} := \left\{f : \left\|f - \hat{f}_{\mathcal{T}^*}\right\|_\infty \leq Q_n(\gamma)\right\}$, where $Q_n(\gamma)$ is chosen such that $\Pi\left[C_n^{L^\infty} | X\right] \geq 1-\gamma$.

Figure 3.4: Posterior sample in the confidence band $\mathcal{C}_n^{\mathcal{M}}$ - $\gamma = 0.05$ and $n = 10^4$.

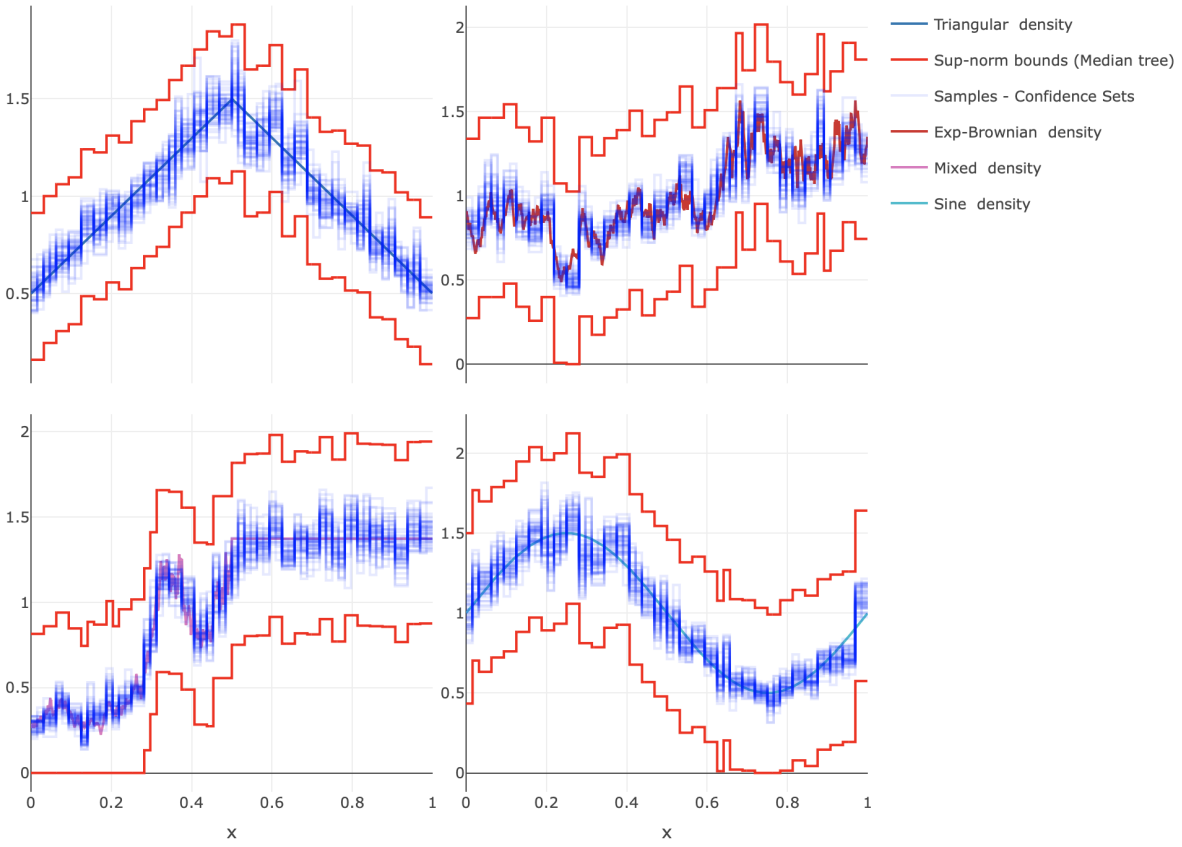
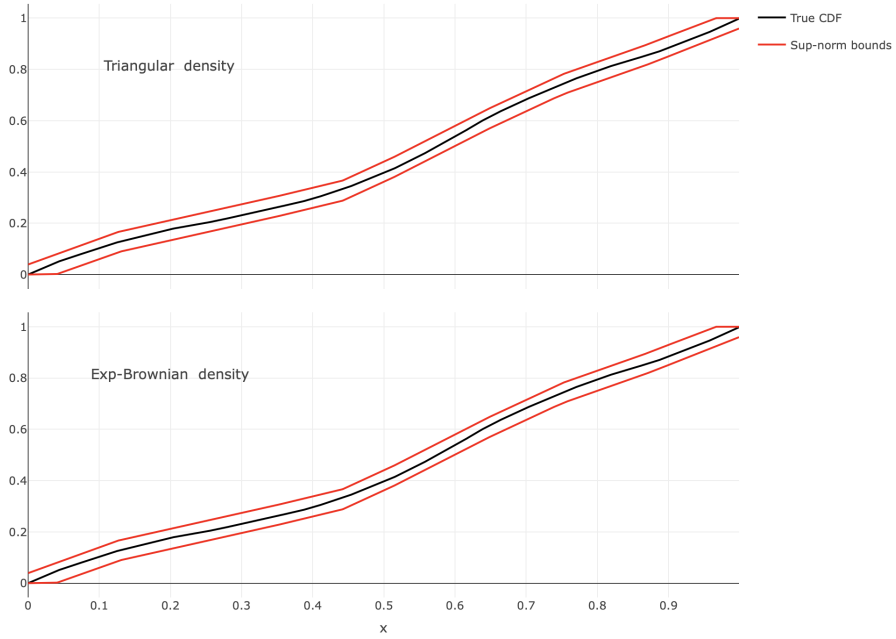


Figure 3.5: Posterior samples in the confidence set $\mathcal{F}_n$ - $n = 10^3$.



In a white noise regression setting, [139] proved that these two sets are asymptotically independent (see Theorem 5.3 therein), in the sense that $\Pi\left[C_n^{L^\infty} \cap \mathcal{C}_n^{\mathcal{M}} | X\right] \overset{P_{f_0}}{\to} (1-\gamma)^2$. As above, we sampled 10^4 draws from the posterior to estimate de posterior credibility of the

different sets, which we present in Table 3.1. The results seem to indicate that the independence phenomenon of the credible sets as described above still hold in the present density estimation setting, as the margin of difference observed is of the order of the Monte-Carlo error. Intuitively speaking, this independence under the posterior if true (at least asymptotically) would mean that the two credible sets reflect *different* aspects of the posterior distribution. Although this result from [139] is seemingly verified in density estimation with an OPT prior, we did not investigate this question from a theoretical point of view in the present chapter; we expect the proof to be significantly more involved than in the (conjugate) Gaussian white noise setting and we leave this point for future work.

3.6 Discussion

In the present work we establish an inference theory for Optional Pólya trees introduced in [173] by deriving posterior contraction rates as well as confidence bands for the problem of uncertainty quantification. By contrast, only posterior consistency had been previously obtained until now for such priors. Although we focus on this class of prior distributions, we point out that our proofs and results also apply to different tree priors, such as ones conditioning on the number of leaves as in Example 4. The results and proofs highlight how beneficial a multiscale approach to study tree-based methods, as introduced in [34], can be.

As for related priors in density estimation, non-adaptive contraction rates were obtained in [29] for Pólya trees for carefully chosen regularity-dependent parameters of the Beta random variables. The addition of a hyperprior on the tree structure in OPTs allows for adaptation, so that the Beta parameters can be set as an arbitrary constant (a similar comment can be done about Spike-and-slab Pólya trees [31]). The Beta variables in the Pólya-like mass allocation mechanism could be replaced by another distribution, but we stuck to them for simplicity of analysis and presentation.

In Section 3.5, we mentioned some further results on OPTs to be investigated. First, tree-based methods have a natural ability to adapt to the local regularity. While this has been proved in [142] in a regression setting, this should also be the case with OPTs in density estimation. Another expected advantage of trees is that in high-dimensional settings, they induce a 'tree–structured' sparsity, which could help in addressing the curse of dimensionality. As original OPTs [114] have been introduced in arbitrary dimensions, it is natural to further our theoretical analysis in this direction, Also, the interesting alleged posterior independence of sets $C_n^{L^\infty}$ and $\mathcal{C}_n^{\mathcal{M}}$ still needs to be proven and would confirm that the two constructions rely on somewhat different aspects of the posterior distribution in density estimation too.

Finally, the almost-optimal rates we obtain are valid for Hölder regularities up to 1. This is related to the fact that samples of OPTs are piecewise constant on some random partition. In order to achieve faster rates for smoother densities, one possibility explored in [107] consists in replacing 'hard' (histogram) trees with 'smooth' trees. Another promising possibility is to look at forests priors. Indeed, the aggregation of many trees tends to result in estimators that are more 'regular' and thereby more suitable to the estimation of smoother objects: for frequentist estimators in regression, this was noted in [3, 121] for regularities $\alpha \leq 2$. The recent work [137] establishes that, when an L^1 or Hellinger loss on densities is considered, forests of Pólya trees enable adaptation to arbitrary regularities α. This will be investigated elsewhere.

3.7 Proof of the main results

Below, the depth $L_n = L_n(\alpha)$ defined as

$$2^{L_n} = c_0(n/\log n)^{\frac{1}{1+2\alpha}}, \tag{3.21}$$

for some $c_0 > 0$, will be helpful in our theoretical analysis. Also, in the below proofs, C stands for a generic constant whose precise value we do not track and can change from line to line.

3.7.1 Proof of Theorem 11

Let's write $T_n = \{\mathcal{T} \,|\, d(\mathcal{T}) \leq L_n,\ S(f_0, \tau) \subset \mathcal{T}\}$, $S(f_0, \tau)$ as in Lemma 22, and $\mathcal{E}_n = \{f : \exists \mathcal{T} \in T_n,\ f$ piecewise constant on $I_{\mathcal{T}}\}$. Moreover, we write, for L_n as in (3.21) and any tree $\mathcal{T} \in \mathbb{T}_n$, the following othogonal projections of f_0: $f_0^{\mathcal{T}}$ onto the span $\{\psi_{lk} \,|\, (l,k) \in \mathcal{T}\}$, $f_0^{L_n^c}$ onto $\{\psi_{lk} \,|\, l > L_n\}$, and $f_0^{\mathcal{T}^c, L_n}$ onto the orthocomplement of the union of the two last spans. For $f_0 \in \mathcal{C}^\alpha[0,1)$, $0 < \alpha \leq 1$, we have in particular that

$$||f_0^{L_n^c}||_\infty \leq \sum_{l > L_n} 2^{l/2} \max_{0 \leq k < 2^l} |f_{0,lk}| \lesssim \sum_{l > L_n} 2^{-l\alpha} \lesssim \left(n^{-1}\log n\right)^{\frac{\alpha}{1+2\alpha}}, \tag{3.22}$$

(see for instance [29]). Then, for any density f_0, we have the upper bound, for $\mathcal{B}_M$ as in Lemma 28,

$$\begin{aligned}
&\Pi\left[\|f - f_0\|_\infty > M_n\epsilon_n \,|\, X\right] \\
&\leq \Pi[\mathcal{E}_n^c \,|\, X]\mathbb{1}_{\mathcal{B}_M} + \Pi\left[\|f - f_0\|_\infty > M_n\epsilon_n,\ f \in \mathcal{E}_n \,|\, X\right]\mathbb{1}_{\mathcal{B}_M} + \mathbb{1}_{\mathcal{B}_M^c}.
\end{aligned}$$

On one hand, Lemma 28 guarantees that $\mathbb{P}_0(\mathcal{B}_M^c) = o(1)$ for M large enough and Lemmas 21 and 22 ensures that

$$E_{f_0}\{\Pi_f[\mathcal{E}_n^c \,|\, X]\mathbb{1}_{\mathcal{B}_M}\} = o(1).$$

On the other hand, we also have the inequality $\|f - f_0\|_\infty \leq \left\|f - f_0^{\mathcal{T}}\right\|_\infty + \left\|f_0^{\mathcal{T}^c, L_n}\right\|_\infty + \left\|f_0^{L_n^c}\right\|_\infty$. This allows us to control the last term in the above upper bound by mean of the Markov inequality:

$$\begin{aligned}
\Pi\Big[f \in \mathcal{E}_n, \quad \|f - f_0\|_\infty > M_n\epsilon_n \,|\, X\Big]\mathbb{1}_{\mathcal{B}_M} &\leq (M_n\epsilon_n)^{-1}\int_{\mathcal{E}_n} \|f - f_0\|_\infty \, d\Pi[f, \mathcal{T} \,|\, X]\mathbb{1}_{\mathcal{B}_M} \\
&\leq (M_n\epsilon_n)^{-1}\Big[\int_{\mathcal{E}_n} \left\|f - f_0^{\mathcal{T}}\right\|_\infty d\Pi[f, \mathcal{T} \,|\, X]\mathbb{1}_{\mathcal{B}_M} \\
&\quad + \int_{\mathcal{E}_n} \left\|f_0^{\mathcal{T}^c, L_n}\right\|_\infty d\Pi[\mathcal{T} \,|\, X]\mathbb{1}_{\mathcal{B}_M} + \left\|f_0^{L_n^c}\right\|_\infty\Big],
\end{aligned} \tag{3.23}$$

and (3.22) ensures that the last term above is $o(1)$. Similarly, for the second term, using the definition of $\mathcal{E}_n$ and denoting L^* the largest integer such that $2^{-L^*(\alpha+1/2)} \geq n^{-1/2}\log n$,

$$\begin{aligned}
||f_0^{\mathcal{T}^c, L_n}||_\infty &\leq \sum_{l \leq L_n} 2^{l/2} \max_{k:(l,k) \notin \mathcal{T}} |f_{0,lk}| \lesssim \sum_{l \leq L_n} 2^{l/2}\left(\max_{0 \leq k < 2^l} |f_{0,lk}| \wedge \log n \,/\, \sqrt{n}\right) \\
&\lesssim \sum_{l \leq L^*} 2^{l/2}\frac{\log n}{\sqrt{n}} + \sum_{L^* < l \leq L_n} 2^{l/2}2^{-l(1/2+\alpha)} \lesssim 2^{L^*/2}\frac{\log n}{\sqrt{n}} + 2^{-L^*\alpha} \lesssim 2^{-L^*\alpha}.
\end{aligned}$$

This allows us to conclude that the second term in the bound (3.23) is also of the order $o(1)$. It remains to bound the first term in the bound that is also of order $o(1)$ according to Lemma 23. This concludes our proof.

It remains to prove the different lemmas we used to upper bound the different terms above.

Lemma 21. *Suppose $f_0 \in \mathcal{F}(\alpha, K, \mu)$, for some $\mu > 0$, $0 < \alpha \leq 1$, $K > 0$, and assume f follows a prior as in Theorem 11. Then, for any $M > 0$ as in Lemma 28 and Γ large enough, on events $\mathcal{B}_M$, we have, as $n \to \infty$,*

$$\Pi[d(\mathcal{T}) > L_n \,|\, X] \to 0,$$

where L_n is as in (3.21).

Proof. Let $\mathcal{T}$ be a tree of depth $L_n < d(\mathcal{T}) = l \leq L_{\max}$. Then, for

$$\tilde{k} = \min_{(2k,l)\in\mathcal{T}} k, \qquad \epsilon = \epsilon\left(\tilde{k}, l-1\right),$$

let $\mathcal{T}^-$ be the corresponding tree whose nodes $(l, 2\tilde{k})$ and $(l, 2\tilde{k}+1)$ have been removed, i.e. $\mathcal{T} = \mathcal{T}^- \cup \left\{(l, 2\tilde{k}), (l, 2\tilde{k}+1)\right\}$. From (3.8) and (3.9), we have

$$\begin{aligned}
\Pi[\mathcal{T} \,|\, X] &= \Pi[\mathcal{T}^- \,|\, X] \frac{p_\epsilon^X}{1-p_\epsilon^X}\left(1-p_{\epsilon 0}^X\right)\left(1-p_{\epsilon 1}^X\right) \\
&= \Pi[\mathcal{T}^- \,|\, X] p_\epsilon \frac{(1-p_{\epsilon 0})(1-p_{\epsilon 1})}{1-p_\epsilon} \nu_\varepsilon^X \\
&\leq \left(1-\Gamma^{-L_n}\right)^{-1} \Pi[\mathcal{T}^- \,|\, X] \frac{2^{N_X(I_\varepsilon)}}{\Gamma^{l+1}} \underbrace{\frac{B(a+N_X(I_{\varepsilon 0}), a+N_X(I_{\varepsilon 1}))}{B(a,a)}}_{=:Q}.
\end{aligned} \tag{3.24}$$

Then, from Lemma 30, we have for $\tilde{n}_0 = N_X(I_{\epsilon 0})$, $\tilde{n}_1 = N_X(I_{\epsilon 1})$ and $\tilde{n} = N_X(I_\epsilon)$, that

$$Q \lesssim \underbrace{\frac{(2a+\tilde{n}_1-1/2)^{\tilde{n}_1}(2a+\tilde{n}_2-1/2)^{\tilde{n}_2}}{(2a+\tilde{n}-1/2)^{\tilde{n}}}}_{=:Q_1} \underbrace{\frac{(2a+\tilde{n}_1-1/2)^{a-1/2}(2a+\tilde{n}_2-1/2)^{a-1/2}}{(2a+\tilde{n}-1/2)^{2a-1/2}}}_{=:Q_2}.$$

Under our assumptions on f_0, on the event $\mathcal{B}_M$ and for n large enough,

$$n_X(I_{l,k}) \geq \frac{\mu}{2} n 2^{-l} \to \infty$$

for any $l \leq L_{\max}$. Under the same conditions,

$$|\tilde{n}_1 - \tilde{n}_2| \leq n\,|P_0(I_{\varepsilon 0}) - P_0(I_{\varepsilon 1})| + 2MM_{n,l} \leq nK2^{-l(1+\alpha)} + 2MM_{n,l}.$$

The last inequality stems from the fact that f_0 is α-Hölder regular. Therefore, on $\mathcal{B}_M$, for n large enough, if we note $v_{\tilde{n}_1,\tilde{n}_2} = \tilde{n}_1 - \tilde{n}_2$, since $\tilde{n} = \tilde{n}_1 + \tilde{n}_2$ and $\log(1+x) \leq x$ for $x > -1$,

$$Q_1 = \exp\left(\tilde{n}_1 \log\left(\frac{1}{2} + \frac{\tilde{n}_1 - \tilde{n}_2 + 2a - 1/2}{2(2a-1/2+\tilde{n})}\right) + \tilde{n}_2 \log\left(\frac{1}{2} - \frac{\tilde{n}_1 - \tilde{n}_2 - 2a + 1/2}{2(2a-1/2+\tilde{n})}\right)\right)$$

$$= \frac{1}{2^{\tilde{n}}} \exp\left(\tilde{n}_1 \log\left(1 + \frac{v_{\tilde{n}_1,\tilde{n}_2} + 2a - 1/2}{2a - 1/2 + \tilde{n}}\right) + \tilde{n}_2 \log\left(1 - \frac{v_{\tilde{n}_1,\tilde{n}_2} - 2a + 1/2}{2a - 1/2 + \tilde{n}}\right)\right)$$
$$\leq \frac{1}{2^{\tilde{n}}} \exp\left(\frac{v^2_{\tilde{n}_1,\tilde{n}_2}}{2a - 1/2 + \tilde{n}} + \frac{\tilde{n}(2a - 1/2)}{2a - 1/2 + \tilde{n}}\right) \leq \frac{C}{2^{\tilde{n}}} \exp\left(\frac{8K^2 n^2 2^{-2l(1+\alpha)}}{\mu n 2^{-l}} + \frac{16M^2 M^2_{n,l}}{\mu n 2^{-l}}\right)$$
$$\leq \frac{C}{2^{\tilde{n}}} \exp\left(\left(8K^2\mu^{-1}c_0^{-1-2\alpha} + 32M^2(\mu \log 2)^{-1}\right) \log n\right).$$

The last inequality stems from $l > L_n$ and the definition of L_n. The last factor is even easier to control as, on $\mathcal{B}_M$,

$$Q_2 \lesssim \left[n2^{-l}\right]^{-1/2} \lesssim n^{-\frac{\alpha}{1+2\alpha}} \log(n)^{-\frac{1/2}{1+2\alpha}}.$$

Finally, this leads us to

$$\Pi[\mathcal{T} \mid X] = o\left(\Pi[\mathcal{T}^- \mid X] \frac{n^{\left(8K^2\mu^{-1}c_0^{-1-2\alpha} + 32M^2(\mu\log 2)^{-1} - \frac{\alpha}{1+2\alpha}\right)}}{\Gamma^l}\right)$$

uniformly on $\mathcal{T}$ such that $L_n < d(\mathcal{T}) = l \leq L_{\max}$. The application $\mathcal{T} \longrightarrow \mathcal{T}^-$ defined above is surjective and is such that each tree $\mathcal{T}^-$ is the image of at most 2^{l-1} trees $\mathcal{T}$. Then, the event of interest verifies for $\Gamma > 2$ and $\bar{C} = 8K^2\mu^{-1}c_0^{-1-2\alpha} + 32M^2(\mu\log 2)^{-1} - \frac{\alpha}{1+2\alpha}$,

$$\Pi[d(\mathcal{T}) > L_n \mid X] = \sum_{l=L_n+1}^{L_{\max}} \Pi[d(\mathcal{T}) = l \mid X] = \sum_{l=L_n+1}^{L_{\max}} \sum_{\mathcal{T}: d(\mathcal{T})=l} \Pi[\mathcal{T} \mid X]$$
$$= o\left(\sum_{l=L_n+1}^{L_{\max}} \sum_{\mathcal{T}: d(\mathcal{T})=l} \Pi[\mathcal{T}^- \mid X] \frac{n^{\bar{C}}}{\Gamma^l}\right) = o\left(\sum_{l=L_n+1}^{L_{\max}} \sum_{\mathcal{T}^-} \Pi[\mathcal{T}^- \mid X] \frac{2^l n^{\bar{C}}}{\Gamma^l}\right) = o\left(\frac{2^{L_n} n^{\bar{C}}}{\Gamma^{L_n}}\right) \tag{3.25}$$

which is $o(1)$ whenever $\{\log\Gamma/\log 2 - 1\}/(1+2\alpha) \geq \bar{C}$, that is, if $\Gamma \geq 2^{1+\bar{C}(1+2\alpha)}$. □

Lemma 22. *Under the same assumptions on f_0 as in Lemma 21, for Π as in Theorem 11 and on the events $\mathcal{B}_M$ from Lemma 28, for $\tau > 0$ large enough and L_n as in (3.21), the set*

$$S(f_0, \tau) := \left\{(l,k) : |f_{0,lk}| \geq \tau \frac{\log n}{\sqrt{n}}\right\}$$

satisfies, as $n \to \infty$,

$$\Pi[\{\mathcal{T} : S(f_0,\tau) \not\subset \mathcal{T}_{int}\} \mid X] \to 0.$$

Proof. First, since $f_0 \in \Sigma(\alpha, K)$ for some $\alpha, K > 0$, there exists $C > 0$ such that, for any $l \geq 0, 0 \leq k < 2^l$, $|f_{0,lk}| \leq C2^{-l(\alpha+1/2)}$. Thus, for τ large enough, $(l,k) \in S(f_0,\tau)$ implies $l \leq L_n$.
Now, let's take (l_S, k_S) a node in $S(f_0,\tau)$. Then, let's define

$$\mathbb{T}_{n,(l_S,k_S)} := \{\mathcal{T} \in \mathbb{T}_n \mid (l_S,k_S) \notin \mathcal{T}_{\text{int}}\},$$

the set of trees in the support of our prior distribution on tree structures that do not have (l_S,k_S) as an internal node, and $\epsilon = \epsilon(k_S,l_S)$. To any tree $\mathcal{T} \in \mathbb{T}_{n,(l_S,k_S)}$, it is possible to

associate the full binary tree $\mathcal{T}^+$ which is the smallest extension of $\mathcal{T}$ with (l_S, k_S) as an interior node,

$$\mathcal{T}^+ = \underset{\mathcal{T}' \in \mathbb{T}_n :\ \mathcal{T} \subset \mathcal{T}',\ (l_S,k_S) \in \mathcal{T}'_{\text{int}}}{\arg\min} |\mathcal{T}'| .$$

This new tree is realized with the completion of the route from the root to the node (l_S, k_S), starting from the leaf node (l_0, k_0) of this route which is included in $\mathcal{T}$. Then, as in (3.24) and using Lemma 30, we now have for some constant $C > 1$,

$$\begin{aligned}
\frac{\Pi[\mathcal{T} \,|\, X]}{\Pi[\mathcal{T}^+ \,|\, X]} &\leq C^{{l_S}^2} \prod_{l=l_0}^{l_S} \left(2^{n_X\left(I_{\epsilon^{[l]}}\right)} \mathrm{B}\Big(a + n_X\left(I_{\epsilon^{[l]}0}\right), a + n_X\left(I_{\epsilon^{[l]}1}\right) \Big) \right)^{-1} \\
&\leq C^{{l_S}^2} \underbrace{\prod_{l=l_0}^{l_S} \frac{(2a + n_X\left(I_{\epsilon^{[l]}}\right) - 1/2)^{2a-1/2}}{(a + n_X\left(I_{\epsilon^{[l]}0}\right) - 1/2)^{a-1/2} (a + n_X\left(I_{\epsilon^{[l]}1}\right) - 1/2)^{a-1/2}}}_{=:\, Q_1} \\
&\qquad \underbrace{\prod_{l=l_0}^{l_S} \frac{(2a + n_X\left(I_{\epsilon^{[l]}}\right) - 1/2)^{n_X\left(I_{\epsilon^{[l]}}\right)}}{2^{n_X\left(I_{\epsilon^{[l]}}\right)} (a + n_X\left(I_{\epsilon^{[l]}0}\right) - 1/2)^{n_X\left(I_{\epsilon^{[l]}0}\right)} (a + n_X\left(I_{\epsilon^{[l]}1}\right) - 1/2)^{n_X\left(I_{\epsilon^{[l]}1}\right)}}}_{=:\, Q_2} .
\end{aligned} \tag{3.26}$$

where we recall that $\epsilon^{[l]}$ denotes the l first elements of the sequence ϵ. On the event $\mathcal{B}_M$, for all $l \leq L_n + 1$ and possible k, we have, using that $f_0 \geq \mu > 0$, $N_X\left(I_{l,k}\right) \gtrsim n 2^{-l} \gtrsim n 2^{-L_n} \to \infty$ as $n \to \infty$. Since it is also upper bounded (as f_0 is a Hölder density), we have $N_X\left(I_{l,k}\right) \lesssim n 2^{-l}$. Therefore, since these bounds are uniform on $l \leq L_n + 1$,

$$Q_1 \leq \prod_{l=l_0}^{l_S} C \left(n 2^{-l} \right)^{1/2} \leq C^{l_S} \sqrt{n}^{l_S}.$$

Also, in Q_2, the factor at index l is equal to, writing $\tilde{n}_0 = N_X\left(I_{\epsilon^{[l]}0}\right), \tilde{n}_1 = N_X\left(I_{\epsilon^{[l]}1}\right), \tilde{n} = N_X\left(I_{\epsilon^{[l]}}\right)$,

$$\exp\left[\tilde{n}_0 \log\left(\frac{2a - 1/2 + \tilde{n}}{2a - 1 + 2\tilde{n}_0} \right) + \tilde{n}_1 \log\left(\frac{2a - 1/2 + \tilde{n}}{2a - 1 + 2\tilde{n}_1} \right) \right].$$

If we write $KL(a; b)$ the Kullback-Leibler divergence between Bernoulli distributions of parameters $0 \leq a, b \leq 1$, then, for n large enough, on $\mathcal{B}_M$, this is bounded by

$$\exp\left[-C\tilde{n} KL\Big(\frac{a - 1/2 + \tilde{n}_0}{2a - 1 + \tilde{n}}; 1/2 \Big) \right] \exp\left[\tilde{n} \log\left(1 + \frac{1}{4a - 2 + 2\tilde{n}} \right) \right].$$

The second factor can be bounded by a constant, uniformly on $l \leq L_n + 1$. The first factor can be bounded by 1 for $l < l_S$, while for $l = l_S$, we can use the bound $KL(a; b) \geq \|\mathrm{Be}(a) - \mathrm{Be}(b)\|_1^2 / 2$ to write

$$\exp\left[-C\tilde{n} KL\Big(\frac{a - 1/2 + \tilde{n}_0}{2a - 1 + \tilde{n}}; 1/2 \Big) \right] \leq \exp\left[-C\tilde{n}^{-1} (\tilde{n}_0 - \tilde{n}_1)^2 \right].$$

By definition $|f_{0,l_S k_S}| = 2^{l_S/2} \left| P_0(I_{(l_S+1)(2k+1)}) - P_0(I_{(l_S+1)(2k)}) \right|$, so that on $\mathcal{B}_M$, $|\tilde{n}_0 - \tilde{n}_1| \geq n\, |f_{0,l_S k_S}|\, 2^{-l_S/2} - 2MM_{n,l_S+1}$, hence the upper bound for τ large enough:

$$\exp\left[-C(\tau \log n - 2M\sqrt{l_S + 1 + L_n})^2 \right] \leq \exp\left[-C\tau^2 \log^2 n \right],$$

where we used the definition of S, M_{n,l_S+1}, L_n and $l_S \leq L_n$.

Finally, for τ large enough and using that $l_S \leq L_n \leq \log n$, we can conclude that there exists constants $C_1, C_2 > 0$ such that

$$\frac{\Pi[\mathcal{T} \,|\, X]}{\Pi[\mathcal{T}^+ \,|\, X]} \leq C_1^{l_S^2} n^{-(C_2\tau^2 - 1/2)\log n} \leq n^{-(C_2\tau^2 - 1/2 - \log C_1)\log n}. \tag{3.27}$$

Since any tree verifying $(l_S, k_S) \in \mathcal{T}$ is the image of at most $l_S + 1$ trees by the map

$$\begin{aligned} \mathbb{T}_{n,(l_S,k_S)} &\to \{\mathcal{T}' \in \mathbb{T}_n : \ (l_S, k_S) \in \mathcal{T}'_{\text{int}}\} \\ \mathcal{T} &\mapsto \mathcal{T}^+ \end{aligned},$$

as it is the length of the path from the root to the node (l_S, k_S) in a tree $\mathcal{T} \in \mathbb{T}_n$,

$$\begin{aligned} \Pi[(l_S, k_S) \notin \mathcal{T} \,|\, X] &= \sum_{\mathcal{T}:(l_S,k_S)\notin\mathcal{T}} \frac{\Pi[\mathcal{T} \,|\, X]}{\Pi[\mathcal{T}^+ \,|\, X]} \Pi[\mathcal{T}^+ \,|\, X] \\ &\leq n^{-(C_2\tau^2 - 1/2 - \log C_1)\log n} (l_S + 1) \sum_{\mathcal{T}:(l_S,k_S)\in\mathcal{T}} \Pi[\mathcal{T} \,|\, X] \\ &\leq n^{-(C_2\tau^2 - 1/2 - \log C_1)\log n} \log n, \end{aligned}$$

which allows us in conjunction with the definition of L_n to conclude that

$$\begin{aligned} \Pi[\{\mathcal{T} : S(f_0, \tau) \not\subset \mathcal{T}\} \,|\, X] &\leq \sum_{(l,k)\in S(f_0,\tau)} \Pi[(l,k) \notin \mathcal{T} \,|\, X] \\ &\leq 2^{L_n+1} n^{-(C_2\tau^2 - 1/2 - \log C_1)\log n} \log n \\ &\to 0 \end{aligned}$$

as $n \to \infty$ for τ large enough. □

Lemma 23. *Let $T_n = \{\mathcal{T} \in \mathbb{T}_n : \ d(\mathcal{T}) \leq L_n, \ S(f_0, \tau) \subset \mathcal{T}\}$ for L_n as in (3.21), $c_0 > 0$ small enough, and $\tau > 0$ as in Lemma 22. Then, under the conditions of Lemma 21 and on the event $\mathcal{B}_M$ for $M > 0$ large enough, there exists a constant $C > 0$ such that for n sufficiently large, uniformly on $\mathcal{T} \in T_n$,*

$$\int \max_{(l,k)\in\mathcal{T}_{int}} |f_{lk} - f_{0,lk}| \, d\Pi[f \,|\, \mathcal{T}, X] \leq C\sqrt{\frac{\log n}{n}}.$$

Proof. Given a tree $\mathcal{T}$, let us define the map $\bar{f}_{\mathcal{T}}$ such that, for each terminal node (l, k) in $\mathcal{T}_{\text{ext}}$ and $x \in I_{lk}$,

$$\bar{f}_{\mathcal{T}}(x) = 2^l \prod_{i=1}^{l} \bar{Y}_{\epsilon[i]}, \quad \epsilon = \epsilon(k, l),$$

where

$$\bar{Y}_\epsilon = E[Y_\epsilon \,|\, X, \mathcal{T}] = \frac{a + N_X(I_{\epsilon 0})}{2a + N_X(I_\epsilon)}.$$

This defines the mean posterior density given the tree structure $\mathcal{T}$. Similarly, for each $(l, k) \in \mathcal{T}$, with $\epsilon = \epsilon(k, l)$, the mean probability measure of I_ϵ is

$$\bar{P}(I_\epsilon) = \prod_{i=1}^{|\epsilon|} \bar{Y}_{\epsilon[i]} =: \bar{p}_\epsilon.$$

Then, expressing the coefficients of the decomposition in the Haar wavelet basis of this mean posterior density, we obtain that for each $(l,k) \in \mathcal{T}_{\text{int}}$, $\epsilon = \epsilon(k,l)$,

$$\bar{f}_{\mathcal{T},lk} \coloneqq \langle \bar{f}_{\mathcal{T}}, \psi_{lk} \rangle = 2^{l/2} \left(\bar{p}_{\epsilon} - 2\bar{p}_{\epsilon 0} \right) = 2^{l/2} \bar{p}_{\epsilon} \left(1 - 2\bar{Y}_{\epsilon 0} \right),$$

while $\bar{f}_{\mathcal{T},lk} = 0$ for $(l,k) \notin \mathcal{T}_{\text{int}}$. When it comes to the true sampling density f_0, we obtain the similar expression, denoting $p_{0,\epsilon} \coloneqq P_0(I_\epsilon)$ and $y_{\epsilon 0} \coloneqq \frac{P_0(I_{\epsilon 0})}{P_0(I_\epsilon)}$,

$$f_{0,lk} = 2^{l/2} p_{0,\epsilon} (1 - 2y_{\epsilon 0}),$$

and, for densities f sampled from the posterior distribution given $\mathcal{T}$, with $p_\epsilon \coloneqq \prod_{i=1}^{|\epsilon|} Y_{\epsilon[i]}$,

$$f_{lk} = 2^{l/2} \tilde{p}_{\epsilon} \left(1 - 2Y_{\epsilon 0} \right) \mathbb{1}_{(l,k) \in \mathcal{T}_{\text{int}}}.$$

From now on, for simplicity of notations, $\epsilon = \epsilon(k,l)$ as the context will make it clear what the pair (l,k) is. For any $\mathcal{T} \in \mathbb{T}_n$, one can bound $|f_{lk} - f_{0,lk}| \leq \left| f_{lk} - \bar{f}_{\mathcal{T},lk} \right| + \left| \bar{f}_{\mathcal{T},lk} - f_{0,lk} \right|$. Using the above expressions, the second term is rewritten as

$$\left| \bar{f}_{\mathcal{T},lk} - f_{0,lk} \right| = \left| f_{0,lk} \left[\frac{\bar{p}_{\epsilon}}{p_{0,\epsilon}} - 1 \right] + 2^{l/2+1} (y_{\epsilon 0} - \bar{Y}_{\epsilon 0}) \right|.$$

Then, as we are on the event $\mathcal{B}_M$, we bound the two terms above by means of Lemmas 1 and 2 from [29] (which are valid for some c_0 small enough) and the bound $p_{0,\epsilon} \lesssim 2^{-|\epsilon|}$ (as f_0 is upper bounded), which give uniformly on $\mathcal{T} \in \mathbb{T}_n$ and $(l,k) \in \mathcal{T}_{\text{int}}$,

$$\begin{aligned} \left| \bar{f}_{\mathcal{T},lk} - f_{0,lk} \right| &\lesssim |f_{0,lk}| \left[a\frac{2^l}{n} + \sqrt{\frac{L_n 2^l}{n}} \right] + \left[|f_{0,lk}| \frac{a2^l}{n} + \sqrt{\frac{L_n}{n}} \right] \\ &\lesssim |f_{0,lk}| \left[a\frac{2^l}{n} + \sqrt{\frac{L_n 2^l}{n}} \right] + \sqrt{\frac{\log n}{n}} \qquad \text{as } L_n \lesssim \log n. \end{aligned} \tag{3.28}$$

Since f_0 is α-Hölder, $|f_{0,lk}| \lesssim 2^{-l(1/2+\alpha)}$, and the last quantity in the above inequality is smaller (up to a constant) than $\sqrt{n^{-1} \log n}$ as $l \leq L_n$. It then remains to bound the term

$$\int \max_{(l,k) \in \mathcal{T}_{\text{int}}} \left| f_{lk} - \bar{f}_{\mathcal{T},lk} \right| d\Pi[f \,|\, \mathcal{T}, X].$$

To do so, let's first define the event

$$\mathcal{A} = \bigcap_{\epsilon : |\epsilon| < L_n} \left\{ |\bar{Y}_{\epsilon 0} - Y_{\epsilon 0}| \leq M' \sqrt{\frac{L_n}{n P_0(I_{\epsilon 0})}} \right\}$$

for $M' > 0$. By Lemma 29, it follows that, for d a small constant,

$$\Pi \left[\mathcal{A}^c \,|\, \mathcal{T}, X \right] \lesssim \sum_{l \leq L_n} 2^l \exp(-CM'^2 \log n) \lesssim 2^{L_n} \exp(-CM'^2 \log n), \tag{3.29}$$

which is smaller than $(n/\log n)^{1/(1+2\alpha)} n^{-CM'^2}$. Then,

$$\left| f_{lk} - \bar{f}_{\mathcal{T},lk} \right| = \left| 2^{l/2+1} \bar{p}_{\epsilon} \left(\bar{Y}_{\epsilon 0} - Y_{\epsilon 0} \right) + \left[\frac{p_{\epsilon}}{\bar{p}_{\epsilon}} - 1 \right] \left(\bar{f}_{\mathcal{T},lk} + 2^{l/2+1} \bar{p}_{\epsilon} (\bar{Y}_{\epsilon 0} - Y_{\epsilon 0}) \right) \right|.$$

Applying Lemmas 2 and 3 from [29] (valid once again for some c_0 small enough), on the events $\mathcal{B}_M$ and $\mathcal{A}$, uniformly on ϵ such that $|\epsilon| = l$ for some $l \leq L_n$,

$$\left|\frac{p_\epsilon}{\bar{p}_\epsilon} - 1\right| \lesssim \sum_{i=1}^{l} \sqrt{\frac{L_n}{nP_0(I_{\epsilon[i]})}} \lesssim \sqrt{\frac{L_n 2^l}{n}}.$$

Therefore, we directly have that on the events $\mathcal{B}_M$ and $\mathcal{A}$,

$$\begin{aligned}\left|f_{lk} - \bar{f}_{\mathcal{T},lk}\right| &\lesssim \left|\bar{f}_{\mathcal{T},lk}\right| \sqrt{\frac{L_n 2^l}{n}} + 2^{l/2}\bar{p}_\epsilon \left[\sqrt{\frac{L_n}{nP_0(I_{\epsilon 0})}} + \frac{L_n}{n}\sqrt{\frac{2^l}{P_0(I_{\epsilon 0})}}\right] \\ &\lesssim \left|\bar{f}_{\mathcal{T},lk}\right| \sqrt{\frac{L_n 2^l}{n}} + \sqrt{\frac{L_n}{n}},\end{aligned} \tag{3.30}$$

where we used that on $\mathcal{B}_M$, $\bar{p}_\epsilon \lesssim 2^{-|\epsilon|}$ for n large enough as f_0 is upper bounded, and $P_0(I_{\epsilon 0}) \gtrsim 2^{-|\epsilon|}$. Finally, with $\left|\bar{f}_{\mathcal{T},lk}\right| \leq \left|\bar{f}_{\mathcal{T},lk} - f_{0,lk}\right| + |f_{0,lk}|$ and using the same computation as for (3.28), we have $\left|f_{lk} - \bar{f}_{\mathcal{T},lk}\right| \lesssim \sqrt{\frac{\log n}{n}}$. This gives

$$\begin{aligned}\int \max_{(l,k)\in\mathcal{T}_{\text{int}}} |f_{lk} - f_{0,lk}|\, d\Pi\left[f \,|\, \mathcal{T}, X\right] &\lesssim \sqrt{\frac{\log n}{n}} + \int_{\mathcal{A}^c} \max_{(l,k)\in\mathcal{T}_{\text{int}}} \left|f_{lk} - \bar{f}_{\mathcal{T},lk}\right| d\Pi[f \,|\, \mathcal{T}, X] \\ &\lesssim \sqrt{\frac{\log n}{n}} + 2^{L_n/2}\Pi[\mathcal{A}^c \,|\, \mathcal{T}, X] \lesssim \sqrt{\frac{\log n}{n}} + \left(\frac{n}{\log n}\right)^{\frac{\alpha/2}{2\alpha+1}} \left(\frac{n}{\log n}\right)^{\frac{1}{1+2\alpha}} n^{-dM'^2} \\ &\lesssim \sqrt{\frac{\log n}{n}} \qquad \text{for } M' \text{ large enough,}\end{aligned} \tag{3.31}$$

where the second inequality comes from the fact that, for a density f, $|\langle f, \psi_{lk}\rangle| \leq 2^{l/2}$. This concludes the proof as this bound holds uniformly on $\mathcal{T} \in T_n$. □

3.7.2 Proofs for confidence bands

Proof of Proposition 12. On the event $\mathcal{E}$ from Lemma 24, the bound on the median tree depth implies that for any $h, g \in C_n$,

$$\begin{aligned}\|h - g\|_\infty &\leq \|h - f_{\mathcal{T}^*}\|_\infty + \|g - f_{\mathcal{T}^*}\|_\infty \\ &\leq 2\sigma_n \\ &\leq 2A^{1/2} v_n \sqrt{\frac{\log n}{n}} 2^{L_n/2} \lesssim v_n \left(\frac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}.\end{aligned}$$

Also, Lemma 27 ensures that

$$\|\hat{f}_{\mathcal{T}^*} - f_0\|_\infty = O_{P_0}\left(\left(\frac{\log^2 n}{n}\right)^{\frac{\alpha}{2\alpha+1}}\right).$$

Then, according to the proof of Proposition 3 in [77], for any $f_0 \in \mathcal{S}(\alpha, K, \eta)$ and l_1 large enough

$$\sup_{(l,k):\ l\geq l_1} |\langle f_0, \psi_{lk}\rangle| \geq C 2^{-l_1(\alpha+1/2)}.$$

For $\Delta_n > 0$ and $\zeta > 0$ such that

$$\zeta \left(\frac{n}{\log^2 n} \right)^{1/(2\alpha+1)} \leq 2^{\Delta_n} \leq 2\zeta \left(\frac{n}{\log^2 n} \right)^{1/(2\alpha+1)},$$

this implies that

$$\sup_{(l,k):\ l \geq \Delta_n} |\langle f_0, \psi_{lk} \rangle| \geq C\zeta^{-\alpha-1/2} \frac{\log n}{\sqrt{n}}.$$

Therefore, if ζ is small enough, there exists $l \geq \Delta_n$ and $0 \leq k < 2^l$ such that $|\langle f_0, \psi_{lk} \rangle| > A \log n / \sqrt{n}$, and then $(l,k) \in \mathcal{T}^*$ on $\mathcal{E}$ according to Lemma 24. As a consequence,

$$\sigma_n \geq v_n \sqrt{\frac{\log n}{n}} 2^{\Delta_n/2} \geq C' \frac{v_n}{\log^{1/2} n} \left(\frac{\log^2 n}{n} \right)^{\alpha/(2\alpha+1)}, \tag{3.32}$$

and since $\log^{1/2} n = o(v_n)$, $\|f_0 - f_{\mathcal{T}^*}\|_\infty \leq \sigma_n/2$ for n large enough. This allows us to conclude that

$$P_0 \left[f_0 \in \mathcal{C}_n \right] = P_0 \left[\{ f_0 \in \mathcal{C}_n \} \cap \mathcal{E} \right] + o(1) = 1 + o(1).$$

It remains to determine the credibility level of the set $\mathcal{C}_n$. From Theorem 11 and Lemma 27, the posterior contracts towards f_0 and the $\hat{f}_{\mathcal{T}^*}$ converges to f_0 on an asymptotically certain event $\mathcal{E}$, both at a faster rate than σ_n (see (3.32)). Therefore, an application of the triangular inequality gives

$$\Pi \left[\mathcal{C}_n \,|\, X \right] \geq \Pi \left[\|f - f_0\|_\infty \leq \sigma_n/2 \,|\, X \right] \mathbb{1}_{\mathcal{E}} + \Pi \left[\mathcal{C}_n \,|\, X \right] \mathbb{1}_{\mathcal{E}^c} = 1 + o_{P_0}(1).$$

□

Proof of Proposition 4. The credibility statement follows from the fact that $\mathcal{C}_n$ (respectively the multiscale ball) has credibility 1 (respectively $1-\gamma$) asymptotically. The diameter statement follows from the inclusion $\mathcal{C}_n^{\mathcal{M}} \subset \mathcal{C}_n$. For coverage, one combines Theorem 12 which gives that $\mathcal{C}_n$ has asymptotic coverage 1, with Theorem 5 in [33] which from the nonparametric BvM (Theorem 14) enables to deduce frequentist coverage of $\|\cdot\|_{\mathcal{M}_0(w)}$–balls (hence the multiscale ball in the intersection defining $\mathcal{C}_n^{\mathcal{M}}$ has asymptotic coverage $1-\gamma$). □

3.8 Supplementary elements

3.8.1 The classical Pólya tree and T–Pólya trees

Let us partition the sample space $I_\varnothing = [0,1)$ as $I_{1,0} \cup I_{1,1}$, these two subsets being the level-1 elementary regions. These can in turn be partitioned as $I_{1,0} = I_{2,0} \cup I_{2,1}$ and $I_{1,1} = I_{2,2} \cup I_{2,3}$, involving level-2 elementary regions. Continuing this partitioning scheme gives the general level-k elementary region, $k \geq 1$, whose set will be written as $\mathcal{A}^k$. More precisely, we partition $I_{l,k} = I_{l+1,2k} \cup I_{l+1,2k+1}$, $l \geq 0, 0 \leq k \leq 2^l - 1$. From this recursive partitioning scheme, one defines a random recursive partition of $I_\varnothing$ and an associated random density.

The Pólya Tree prior corresponding to the partitioning $\cup_{l=1}^{\infty} \mathcal{A}^l$ is the distribution on probability measure on $[0;1)$, whose samples are defined by the conditional probabilities

$$\epsilon \in \mathcal{E}^*, \ P\left(I_{\epsilon 0} | I_\epsilon\right) = V_{\epsilon 0} \sim \text{Beta}(\nu_{\epsilon 0}, \nu_{\epsilon 0}). \tag{3.33}$$

For an appropriate choice of Beta parameters $\nu_\epsilon,\ \epsilon \in \mathcal{E}^*$, samples from this prior actually extends almost surely to an absolutely continuous measure, so that it can be seen as a prior on densities. The Beta random variables $V_{\epsilon 0}$ then corresponds to the share of the mass on I_ϵ that is allocated to $I_{\epsilon 0}$. This mass allocation scheme is illustrated on Figure 3.6: the random mass of each interval I_ϵ is the product of Beta variables on the edges of the path from the root to the corresponding node. As a consequence, the random mass on I_ϵ, $\epsilon \in \mathcal{E}^*$, is equal to $\prod_{i=1}^{|\epsilon|} V_{\epsilon[i]}$.

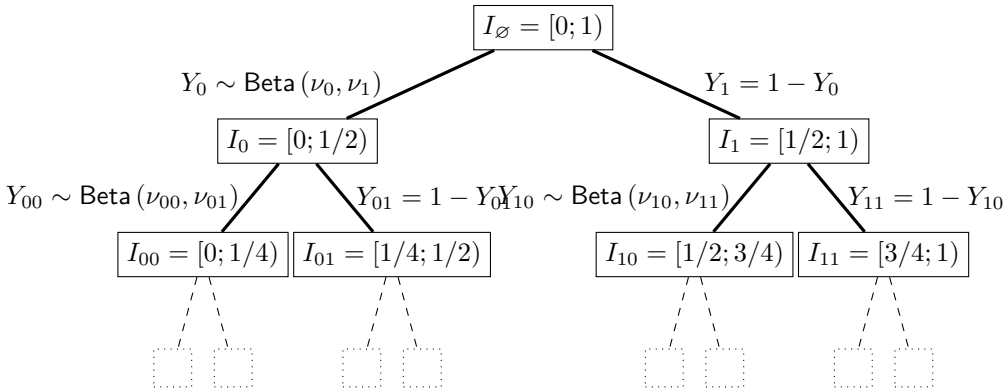


Figure 3.6: Pólya Tree process on the dyadic recursive partitioning, with splits at midpoints.

A simpler related prior on densities, the truncated Pólya Tree prior, stops the splitting of the mass at some level $L < \infty$ and has sampled densities which are constant on each set I_ϵ in $\mathcal{A}^L$, with value $\mu(I_\epsilon)^{-1} \prod_{i=1}^{|\epsilon|} V_{\epsilon[i]}$. If one introduces the tree T as

$$T = \left\{(k,l),\ l \leq L, 0 \leq k < 2^l\right\},$$

that is the complete binary tree of depth $d(T) = L$, it corresponds to a T-Pólya tree distribution with $\Pi_{\mathbb{T}} = \delta_T$.

3.8.2 Tree posteriors: the Galton–Watson/Pólya tree case

As shown in Subsection 3.2.3, the Markov process on trees $GW(p)$ can be seen as a distribution on partitions. We first show that it corresponds to the distribution introduced in [173].

In the Optional Pólya Tree (OPT) construction, different recursive partitioning mechanism are allowed: each level-k elementary region $A \in \mathcal{A}^k$ can be split in $M(A)$ different ways, the j-th being written as

$$A = \cup_{i=1}^{K_j(A)} A_k^j, \tag{3.34}$$

where the A_k^j are level-$(k+1)$ elementary regions (see Appendix 3.8.1). Then, a random partition of the sample space $[0;1)$ is produced recursively. For $0 \leq \rho([0,1)) \leq 1$, the partition is the sample space itself with probability $\rho([0,1))$. Otherwise, one of the $M([0,1))$ partitions are drawn according to probability vector $\lambda([0;1)) = \left(\lambda_1, \ldots, \lambda_{M([0,1))}\right)$. The partitioning then continues: each elementary region A stays intact with probability $\rho(A)$, otherwise it is split (a decision encoded by the variable $S(A) \sim \text{B}(\rho(A))$) and its partition is chosen according to probability vector $\lambda(A)$. Following the discussion in Subsection 3.2.3, the $GW(p)$ is a particular case, where $M(A) = 1$, $\lambda(A) = 1$ and $K_1(A) = 2$, as the intervals are only ever split at their midpoints,

$$I_{l,k} = I_{l+1,2k} \cup I_{l+1,2k+1}. \tag{3.35}$$

The level-k elementary regions are the I_ϵ with $|\epsilon| = k$. Also, it corresponds to the choice of

$$\rho(I_{l,k}) = 1 - p_{lk},\ l < L_{\text{max}}, \qquad \rho(I_{L_{\text{max}},k}) = 0.$$

Given a partition $\mathcal{I}$, in OPT, a probability measure Q is defined by the conditional probabilities, for A an elementary region split as in (3.34),

$$\left(Q(A_1^j|A), \ldots, Q(A_{K_j(A)}^j|A)\right) = Q(A)\theta(A), \qquad \theta(A) \sim \text{Dir}\left(\alpha_1^j(A), \ldots, \alpha_{K_j(A)}^j(A)\right),$$

with Dirichlet random variables θ mutually independent and independent from the variables $S(A')$ for $A \not\subset A'$, and $Q([0,1)) = 1$. For $M(A) = 1$ and $K_1(A) = 2$, it is similar to the mass allocation mechanism in (3.33) when $\alpha_1^1 = \alpha_2^1 = a$. However, whenever the recursive partitioning stops and gives a finite partition, these equations do not completely characterize a measure on Borelians of $[0,1)$, so that the measure Q is defined on Borelians B as

$$Q(B) = \sum_{A \in \mathcal{I}} Q(A) \frac{\mu(A \cap B)}{\mu(A)}.$$

This corresponds to the absolutely continuous measure with density constant on the elements of $\mathcal{I}$. Therefore, the distribution from Proposition 1 is actually a special case of OPT.

3.8.3 The OPT posterior on trees

In the following, we prove Propositions 2 and 3. We first obtain a general formula for the posterior on trees, which implies an explicit formulation of $\Pi[\cdot \mid X, \mathcal{T}]$, and then focus on the OPT prior. The posterior distribution on trees is given for $T \in \mathbb{T}_n$ by Bayes' formula as

$$\Pi[T|X] = \frac{\int \Pi[X, T|f]\, d\Pi[f]}{\int \Pi[X|f]\, d\Pi[f]}.$$

Since $\Pi[X, T|f] = \mathbb{1}_{\mathcal{T}=T} \prod_{i=1}^n f(X_i)$, the numerator is equal to

$$\sum_{T' \in \mathbb{T}_n} \Pi[\mathcal{T} = T']\, \mathbb{1}_{\mathcal{T}=T} \int \prod_{i=1}^n f(X_i)\, d\Pi[f|T'] = \Pi[\mathcal{T} = T] \int \prod_{i=1}^n f(X_i)\, d\Pi[f|T].$$

Writing $N_T(X) := \int \prod_{i=1}^n f(X_i)\, d\Pi[f|T]$ the marginal likelihood, the denominator can be expressed as

$$\sum_{T' \in \mathbb{T}_n} \Pi[\mathcal{T} = T'] \int \Pi[X, \mathcal{T} = T'|f]\, d\Pi[f] = \sum_{T' \in \mathbb{T}_n} \Pi[\mathcal{T} = T']\, N_{T'}(X).$$

Let's compute $N_T(X)$. By definition, for any $i = 1, \ldots, n$,

$$f(X_i) = \prod_{(l,k) \in T_{\text{ext}}} \left(\prod_{j=1}^l 2Y_{\epsilon(k,l)[j]} \right)^{\mathbb{1}_{X_i \in I_{lk}}},$$

and

$$\prod_{i=1}^n f(X_i) = \prod_{(l,k) \in T_{\text{ext}}} \left(\prod_{j=1}^l 2Y_{\epsilon(k,l)[j]} \right)^{N_X(I_{lk})}$$

$$= \prod_{(l,k)\in T\setminus\{(0,0)\}} \left(2Y_{\epsilon(k,l)}\right)^{N_X(I_{lk})}$$
$$= \prod_{(l,k)\in T_{\text{int}}} \left(2Y_{\epsilon(k,l)0}\right)^{N_X\left(I_{\epsilon(k,l)0}\right)} \left(2(1-Y_{\epsilon(k,l)0})\right)^{N_X\left(I_{\epsilon(k,l)1}\right)}.$$

On the one hand, we obtain that

$$\begin{aligned}\Pi[f \mid X, \mathcal{T}] &= N_T(X)^{-1}\Pi[f, X \mid \mathcal{T}] = N_T(X)^{-1}\Pi[X \mid f, \mathcal{T}]\Pi[f \mid \mathcal{T}] \\ &= C(X,T)\prod_{i=1}^{n} f(X_i) \prod_{(l,k)\in T_{\text{ext}}} \prod_{j=1}^{l} Y_{\epsilon(k,l)^{[j]}}^{a}\left(1-Y_{\epsilon(k,l)^{[j]}}\right)^{a} \\ &= C(X,T) \prod_{(l,k)\in T_{\text{int}}} Y_{\epsilon(k,l)0}^{a+N_X\left(I_{\epsilon(k,l)0}\right)} \left((1-Y_{\epsilon(k,l)0})\right)^{a+N_X\left(I_{\epsilon(k,l)1}\right)},\end{aligned}$$

for $C(X,T)$ a constant depending on X and T only, which proves the claim of Proposition 2. On the other hand, for any variable $Y \sim \text{Beta}(a,a)$, one obtains

$$E\left[Y^N(1-Y)^M\right] = \int_0^1 y^N(1-y)^M \frac{y^a(1-y)^a}{B(a.a)}dy = \frac{B(a+N, a+M)}{B(a,a)}.$$

Therefore,

$$N_T(X) = \prod_{(l,k)\in T_{\text{int}}} 2^{N_X\left(I_{\epsilon(k,l)}\right)} \frac{B\left(a+N_X\left(I_{\epsilon(k,l)0}\right), a+N_X\left(I_{\epsilon(k,l)1}\right)\right)}{B(a,a)}.$$

Let's now focus on the special case of the $\text{GW}(p)$ tree prior, as in Proposition 3. For any possible pair (l,k), take $T \in \mathbb{T}_n$ such that $(l,k) \in T_{\text{ext}}$ and let

$$T^+ = T \cup \{(l+1, 2k), (l+1, 2k+1)\}.$$

Then,

$$\Pi[T^+] = \Pi[T]\frac{p_{lk}}{1-p_{lk}}(1-p_{l+1,2k})(1-p_{l+1,2k+1}), \tag{3.36}$$

and

$$\begin{aligned}\frac{\Pi[T^+|X]}{\Pi[T|X]} &= \frac{\Pi\left[\mathcal{T}=T^+\right]L_{T^+}(X)}{\Pi\left[\mathcal{T}=T\right]L_T(X)} \\ &= \frac{p_{lk}}{1-p_{lk}}(1-p_{l+1,2k})(1-p_{l+1,2k+1}) \\ &\quad 2^{N_X\left(I_{\epsilon(k,l)}\right)}\frac{B\left(a+N_X\left(I_{\epsilon(k,l)0}\right), a+N_X\left(I_{\epsilon(k,l)1}\right)\right)}{B(a,a)}.\end{aligned} \tag{3.37}$$

This last quantity is independent of T and T^+ and depends only on (l,k). Therefore, if we can find $p^X_{lk}, p^X_{l+1,2k}, p^X_{l+1,2k+1}$ such that the last quantity in (3.37) is equal to

$$\frac{p^X_{lk}}{1-p^X_{lk}}(1-p^X_{l+1,2k})(1-p^X_{l+1,2k+1}),$$

for any appropriate (l,k), we obtain a formula similar to (3.36) and the posterior on trees is a $GW(p^X)$ process. This defines a set of equations that has a solution, as for any $0 \le k < 2^{L_{\text{max}}}$, we necessarily have $p_{L_{\text{max}}k} = 0$ and the equations can be solved to obtain p^X, starting from $l = L_{\text{max}}$ and solving the successive equations in a "bottom–up" way up to the level $l = 0$.

3.8.4 Median tree properties

Lemma 24. *Under the same prior and assumptions as in Theorem 11, there exists an event $\mathcal{E}$, such that $P_0[\mathcal{E}] = 1 + o(1)$, on which the following is true: for some constants $A > 0, B > 0$,*

- $2^{d(\mathcal{T}^*)} \leq A 2^{L_n} \asymp (n/\log n)^{1/(2\alpha+1)}$, *$L_n$ as in (3.21),*
- *For any (l,k) such that $|f_{0,lk}| \geq Bn^{-1/2}\log n$, $(l,k) \in \mathcal{T}^*_{int}$.*

Proof. On the event $\mathcal{B}_M$ from Lemma 28, Lemma 22 shows that the set $\mathbb{T}^{(2)}$ of trees satisfying the second condition in the lemma, for B large enough, is such that $\Pi\left[\mathbb{T}^{(2)} | X\right] \to 1$. Therefore the event

$$\tilde{\mathcal{E}} = \left\{\Pi\left[\mathbb{T}^{(2)} | X\right] \geq 3/4\right\} \supset \mathcal{B}_M$$

is asymptotically certain.
For any node (l,k) such that $|f_{0,lk}| \geq Bn^{-1/2}\log n$, since it belongs to the interior nodes of any tree in $\mathbb{T}^{(2)}$ by definition,

$$\Pi[(l,k) \in \mathcal{T}_{\text{int}} | X] = \sum_{\mathcal{T} \in \mathbb{T}_n :\ (l,k) \in \mathcal{T}_{\text{int}}} \Pi[\mathcal{T} | X] \geq \Pi\left[\mathbb{T}^{(2)} | X\right].$$

Then, on $\tilde{\mathcal{E}}$, $(l,k) \in \mathcal{T}^*$ by definition and $\mathcal{T}^*$ satisfies the second condition of the lemma.
Let's now turn to the set $\mathbb{T}^{(1)}$ of trees satisfying the first condition in the lemma. Using the same arguments as for (3.25), there exists $C > 0$ such that for any l such that $2^l \gtrsim 2^{L_n}$ and $\Gamma > 0$ large enough,

$$\Pi\left[d\left(\mathcal{T}\right) > l \mid X\right] \leq n^C (2/\Gamma)^l,$$

which holds on the event $\mathcal{B}_M$. Then, since

$$\Pi[(l,k) \in \mathcal{T}_{\text{int}} | X] \leq \Pi[d(\mathcal{T}) > l | X],$$

Markov's inequality implies

$$\begin{aligned}
P_0\left[\left\{\mathcal{T}^* \notin \mathbb{T}^{(1)}\right\} \cap \mathcal{B}_M\right] &= P_0\left[\left\{\exists (l,k) :\ 2^l > A2^{L_n},\ (l,k) \in \mathcal{T}^*\right\} \cap \mathcal{B}_M\right] \\
&\leq \sum_{l:\ 2^l > A2^{L_n}}^{L_{\max}} \sum_{0 \leq k < 2^l - 1} P_0\left[\left\{\Pi[(l-1, \lfloor k/2 \rfloor) \in \mathcal{T}_{\text{int}} \mid X] > 1/2\right\} \cap \mathcal{B}_M\right] \\
&\leq \sum_{l:\ 2^l > A2^{L_n}}^{L_{\max}} 2 \sum_{0 \leq k < 2^l - 1} E_0\left[\Pi[(l-1, \lfloor k/2 \rfloor) \in \mathcal{T}_{\text{int}} \mid X] \mathbb{1}_{\mathcal{B}_M}\right] \\
&= o(1) \text{ for } \Gamma \text{ large enough.}
\end{aligned}$$

One concludes by noting that $\mathcal{B}_M$ is asymptotically certain according to Lemma 28, and $\mathcal{E} = \left\{\mathcal{T}^* \in \mathbb{T}^{(1)}\right\} \cap \mathcal{B}_M$ satisfies the conditions of the lemma. □

Lemma 25. *Let $0 < \alpha \leq 1$, $K > 0$, $\mu > 0$ and $\eta > 0$. Let Π be the same prior as in Theorem 12, then for $f_0 \in \mathcal{S}(\alpha, K, \eta) \cap \mathcal{F}(\alpha, K, \mu)$,*

$$\left(n/\log^2 n\right)^{1/(2\alpha+1)} \lesssim 2^{d(\mathcal{T}^*)} \lesssim (n/\log n)^{1/(2\alpha+1)},$$

on an event of probability converging to 1.

Proof. Using the same argument as above (3.32), we obtain the lower bound. Lemma 24 gives the upper bound. □

Lemma 26. *Let f_0 and ℓ_0 be as in Theorem 14, Π as in Proposition 24 and $\hat{f}_{\mathcal{T}^*}$ as defined in (3.14). The median tree estimator then satisfies*

$$\max_{l>\ell_0(n)} \max_k |\hat{f}_{\mathcal{T}^*,lk} - f_{0,lk}| = O_{P_0}\left(\frac{\log n}{\sqrt{n}}\right).$$

Proof. Let $Q = \max_{l>\ell_0(n)} \max_k |\hat{f}_{\mathcal{T}^*,lk} - f_{0,lk}|$. On the event $\mathcal{E}$ from Proposition 24, one has for B as in the proposition,

$$Q \leq \left(B\frac{\log n}{n^{1/2}}\right) \vee \max_{(l,k)\in\mathcal{T}^*_{\text{int}}, l>\ell_0(n)} |\hat{f}_{\mathcal{T}^*,lk} - f_{0,lk}|.$$

Indeed, for $(l,k) \notin \mathcal{T}^*_{\text{int}}$, we necessarily have $\hat{f}_{\mathcal{T}^*,lk} = 0$ and $|f_{0,lk}| < Bn^{-1/2}\log n$ on $\mathcal{E}$. From (3.41), it also follows that for A as in the proposition and L_n defined in (3.21)

$$\max_{(l,k)\in\mathcal{T}^*_{\text{int}},\ l>\ell_0(n)} |\hat{f}_{\mathcal{T}^*,lk} - f_{0,lk}| \leq \max_{(l,k),\ 2^{\ell_0(n)}<2^l<A2^{L_n}} |P_n\psi_{lk} - P_0\psi_{lk}| =: Q_n.$$

We have that

$$|P_n\psi_{lk} - P_0\psi_{lk}| \leq 2^{l/2}n^{-1}\left(|N(I_{l+1,2k}) - nP_0(I_{l+1,2k})| + |N(I_{l+1,2k+1}) - nP_0(I_{l+1,2k+1})|\right).$$

Therefore, on the event $\mathcal{B}_M$ from Lemma 28, for some constant C depending on M, A, c_0 and α only, and any l as in the above supremum,

$$|P_n\psi_{lk} - P_0\psi_{lk}| \leq C\sqrt{\frac{\log n}{n}}. \tag{3.38}$$

It follows that $Q \lesssim n^{-1/2}\log n$ on the event $\mathcal{E}\cap\mathcal{B}_M$ that is such that $P_0(\mathcal{E}\cap\mathcal{B}_M) = 1 + o(1)$. □

Lemma 27. *Let $\mathcal{T}^*$ as in (3.13) and $\hat{f}_{\mathcal{T}^*}$ as in (3.14). Then, for $f_0 \in \mathcal{F}(\alpha, K, \mu)$,*

$$\|\hat{f}_{\mathcal{T}^*} - f_0\|_\infty = O_{P_0}\left(\left(\frac{\log^2 n}{n}\right)^{\frac{\alpha}{2\alpha+1}}\right).$$

Proof. Let $\mathcal{E}$ as in Lemma 24 and $\mathcal{B}_M$ as in Lemma 28. On $\mathcal{E}\cap\mathcal{B}_M$, for M large enough,

$$\begin{aligned}\|f_0 - f_{\mathcal{T}^*}\|_\infty \leq &\sum_{l:\ 2^l<A2^{L_n}} 2^{l/2}\max\left[\max_{0\leq k<2^l,\ (l,k)\in\mathcal{T}^*_{\text{int}}} |\langle f_0 - f_{\mathcal{T}^*}, \psi_{lk}\rangle|,\ \max_{0\leq k<2^l,\ (l,k)\notin\mathcal{T}^*_{\text{int}}} |\langle f_0, \psi_{lk}\rangle|\right] \\ &+ \sum_{l:\ 2^l\geq A2^{L_n}} 2^{l/2}\max_{0\leq k<2^l} |\langle f_0, \psi_{lk}\rangle|,\end{aligned}$$

using the usual inequality for densities h, g, $\|h-g\|_\infty \leq \sum_{l\geq 0} 2^{l/2}\max_{0\leq k<2^l} |\langle h-g, \psi_{lk}\rangle|$. Since $f_0 \in \Sigma(\alpha, K)$, the second term is smaller than $2^{-\alpha L_n} = O\left((n/\log n)^{-\alpha/(2\alpha+1)}\right)$ (up to a constant depending only on α, K and the constant A from Lemma 24). Then, the first term can itself be upper bounded by the sum of

$$\sum_{l:\ 2^l<A2^{L_n}} 2^{l/2}\max_{0\leq k<2^l,\ (l,k)\in\mathcal{T}^*_{\text{int}}} |\langle f_0 - f_{\mathcal{T}^*}, \psi_{lk}\rangle| \lesssim 2^{L_n/2}\sqrt{\frac{\log n}{n}} = o\left(\left(\frac{\log^2 n}{n}\right)^{\alpha/(2\alpha+1)}\right),$$

where we used that the argument of 3.38 can be extended to $l \leq \ell_0(n)$ on $\mathcal{E} \cap \mathcal{B}_M$, and the term

$$\sum_{l:\ 2^l < A2^{L_n}} 2^{l/2} \max_{0 \leq k < 2^l,\ (l,k) \notin \mathcal{T}^*_{\text{int}}} |\langle f_0, \psi_{lk} \rangle|.$$

It remains to upper bound this last quantity. Let's introduce

$$L^* = \max \left\{ l : \max_{0 \leq k < 2^l} |\langle f_0, \psi_{lk} \rangle| \geq B n^{-1/2} \log n \right\}$$

which is such that $2^{L^*} \asymp \left(\frac{n^{1/2}}{\log n} \right)^{1/(\alpha+1/2)}$ since $\max_{0 \leq k < 2^l} |\langle f_0, \psi_{lk} \rangle| \lesssim 2^{-l(1/2+\alpha)}$. Then, on the event $\mathcal{E}$, the term in the above display is bounded by

$$\begin{aligned} \sum_{l:\ 2^l < A2^{L_n}} 2^{l/2} \left(B \frac{\log n}{\sqrt{n}} \right) \wedge \max_{0 \leq k < 2^l} |\langle f_0, \psi_{lk} \rangle| &\leq \sum_{l:\ l \leq L^*} 2^{l/2} \left(B \frac{\log n}{\sqrt{n}} \right) + \\ &\quad \sum_{l:\ 2^{L^*} < 2^l < A2^{L_n}} 2^{l/2} \max_{0 \leq k < 2^l} |\langle f_0, \psi_{lk} \rangle| \\ &\lesssim \sqrt{2^{L^*} \frac{\log^2 n}{n}} + 2^{-\alpha L^*} \lesssim \left(\frac{\log^2 n}{n} \right)^{\alpha/(2\alpha+1)}. \end{aligned}$$

Combining the previous bounds leads to, on $\mathcal{E} \cap \mathcal{B}_M$,

$$\|f_0 - f_{\mathcal{T}^*}\|_\infty \leq C \left(\log^2 n / n \right)^{\alpha/(2\alpha+1)}.$$

□

3.8.5 Nonparametric BvM theorem

Space $\mathcal{M}_0$ and limiting Gaussian process $\mathcal{N}$

Recall the definition of the space $\mathcal{M}_0$ from (3.18), using an 'admissible' sequence $w = (w_l)_{l \geq 0}$ such that $w_l / \sqrt{l} \to \infty$ as $l \to \infty$,

$$\mathcal{M}_0 = \mathcal{M}_0(w) = \left\{ x = (x_{lk})_{l,k} \ ; \ \lim_{l \to \infty} \max_{0 \leq k < 2^l} \frac{|x_{lk}|}{w_l} = 0 \right\}.$$

Equipped with the norm $\|x\|_{\mathcal{M}_0} = \sup_{l \geq 0} \max_{0 \leq k < 2^l} |x_{lk}| / w_l$, this is a separable Banach space. In a slight abuse of notation, we write $f \in \mathcal{M}_0$ if the sequence of its Haar wavelet coefficients belongs to that space $(\langle f, \psi_{lk} \rangle)_{l,k} \in \mathcal{M}_0$ and for a process $(Z(f),\ f \in L^2)$, we write $Z \in \mathcal{M}_0$ if the sequence $(Z(\psi_{lk}))_{l,k}$ belongs to $\mathcal{M}_0(w)$ almost surely.

White bridge process. For P a probability distribution on $[0,1]$, following [33] one defines the P-white bridge process, denoted by $\mathbb{G}_P$, as the centered Gaussian process indexed by the Hilbert space $L^2(P) = \{ f : [0,1] \to \mathbb{R}; \int_0^1 f^2 dP < \infty \}$ with covariance

$$E[\mathbb{G}_P(f) \mathbb{G}_P(g)] = \int_0^1 (f - \int_0^1 f dP)(g - \int_0^1 g dP) dP. \tag{3.39}$$

We denote by $\mathcal{N}$ the law induced by $\mathbb{G}_{P_0}$ (with $P_0 = P_{f_0}$) on $\mathcal{M}_0(w)$. The sequence $(\mathbb{G}_P(\psi_{lk}))_{l,k}$ indeed defines a tight Borel Gaussian variable in $\mathcal{M}_0(w)$, by Remark 1 of [33].

Admissible sequences (w_l). The main purpose of the sequence (w_l) is to ensure that $(\mathbb{G}_P(\psi_{lk}))_{l,k}$ belongs to $\mathcal{M}_0$. We refer to [33], Section 2.1 and Remark 1, for more background on the choice of (w_l) in the present multiscale setting, and to [32], Section 1.2, for a similar discussion in an Hilbert space setting where the targeted loss is the L^2–norm.

To establish a nonparametric Bernstein–von Mises (BvM) result, following [33] one first finds a space $\mathcal{M}_0$ large enough to have convergence at rate $\sqrt{n}$ of the posterior density to a Gaussian process. One can then derive results for some other spaces $\mathcal{F}$ using continuous mapping for continuous functionals $\psi : \mathcal{M}_0 \to \mathcal{F}$.

Recentering the distribution. To establish the BvM result, one also has to find a suitable way to center the posterior distribution. A possible centering is the median tree estimator $\hat{f}_{\mathcal{T}^*}$ as in (3.14). Other centerings are possible, typically appropriately 'smoothed' versions of the empirical measure P_n associated to the sample $X_1, \dots, X_n$

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}. \tag{3.40}$$

Let us now also note that another way to write the median tree estimator (3.14) is

$$f_{\mathcal{T}^*} = 1 + \sum_{(l,k)\in\mathcal{T}^*_{\text{int}}} (P_n\psi_{lk}) \cdot \psi_{lk}, \tag{3.41}$$

where $P_n\psi_{lk} = n^{-1}\sum_{i=1}^n \psi_{lk}(X_i)$ are the empirical wavelet coefficients, and only terms corresponding to interior nodes (l,k) in the median tree $\mathcal{T}^*$ are active in the sum from the last display. From this we see that the median tree estimator (3.14) can also be interpreted as a smoothed (or 'truncated') version of the empirical measure P_n in (3.40), with truncation occuring along the median tree $\mathcal{T}^*$. Note also that if the prior Π has flat initialisation up to level $l_0(n)$, then all nodes (l,k) with $l \le l_0(n)$ are present in the above sum over $(l,k) \in \mathcal{T}^*_{\text{int}}$.

Nonparametric BvM: statement

For the following result, we work with OPTs with flat initialisation as defined in Section 3.4.3. This is discussed below the next statement.

We have the following Bernstein-von Mises phenomenon for f_0 in Hölder balls. For C_n a function to be specified, we denote by τ_{C_n} the map $\tau_{C_n} : f \to \sqrt{n}(f - C_n)$.

Theorem 14. *Let $\mathcal{N}$ denote the distribution induced on $\mathcal{M}_0(w)$ by the P_0–white bridge $\mathbb{G}_{P_0}$ as defined in (3.39) and let $C_n = \hat{f}_{\mathcal{T}^*}$ the median tree estimator as in (3.14). Let Π be an OPT prior with flat initialisation with $l_0(n)$ that verifies $\sqrt{\log n} \le l_0(n) \le \log n/\log\log n$, and other than that for $l > l_0(n)$ with same parameters as the prior in Theorem 11. Then for every $\alpha \in (0,1]$, for $\mu > 0$, $K \ge 0$ and $\eta > 0$,*

$$\sup_{f_0\in\mathcal{F}(\alpha,K,\mu)} E_{f_0}\left[\beta_{\mathcal{M}_0(w)}(\Pi(\cdot|X)\circ\tau_{C_n}^{-1}, \mathcal{N})\right] \to 0,$$

as $n \to \infty$, for the admissible sequence $w_l = l^{2+\delta}$ for some $\delta > 0$.

Remark 1. *Recalling that the typical nonparametric cut–off sequence $\mathcal{L}$ verifies $2^{\mathcal{L}} \asymp n^{1/(1+2\alpha)}$, assuming $\ell_0(n) = o(\log n)$ amounts to say that $\ell_0(n)$ does not 'interfere' with the nonparametric cut-off $\mathcal{L}$. Similar choices are made in [139], Corollary 3.6. Other choices of sequence $\ell_0(n)$ would also be possible, up to adjusting the sequence (w_l) – one can check that it suffices to have an increasing sequence (w_l) such that $w_{l_0(n)}/\log n \to \infty$ (see, e.g. Theorem S–3 in the Supplement of [34]) –; we do not consider these refinements here.*

Theorem 14 states that the posterior limiting distribution is Gaussian after rescaling; note that, similar to the first such result recently obtained in [139], one slightly modifies the OPT prior to fit the first levels by assuming a flat initialisation. This is in fact necessary for the result to hold, as otherwise the posterior would not be tight at rate $1/\sqrt{n}$ in the space $\mathcal{M}_0(w)$, as was noted in the white noise model in [139], Proposition 3.7. Let us also briefly comment on the recentering C_n: as follows from the proof of Theorem 14, one can replace $C_n = \hat{f}_{\mathcal{T}^*}$ by another estimator that fits all first wavelet coefficients up to $\ell_0(n)$ and such that $\|C_n - f_0\|_{\mathcal{M}_0(\bar{w})} = O_{P_0}(1/\sqrt{n})$, for $\bar{w}$ as in that proof, see also Remark 2 for more on this.

Nonparametric BvM: implications

Using the methods of [33], this result leads to several applications. A first direct implication (this follows from Theorem 5 in [33]) is the derivation of a confidence set in $\mathcal{M}_0(w)$. Setting

$$\mathcal{D}_n = \left\{ f = (f_{lk}) : \ \|f - C_n\|_{\mathcal{M}_0(w)} \leq \frac{R_n}{\sqrt{n}} \right\}, \tag{3.42}$$

where R_n is chosen in such a way that $\Pi[\mathcal{D}_n \,|\, X] = 1 - \gamma$, for some $\gamma > 0$ (or taking the generalised quantile for the posterior radius if the equation has no solution) leads to a set $\mathcal{D}_n$ with the following properties: it is a credible set by definition which is also asymptotically a confidence set in $\mathcal{M}_0(w)$ and the rescaled radius R_n is bounded in probability. Other applications are BvM theorems for functionals, as given a continuous map $\psi : \mathcal{M}_0(w) \to \mathcal{E}$ for some metric space $\mathcal{E}$, convergence results in $\mathcal{M}_0(w)$ can be translated into convergence in $\mathcal{E}$ via the continuous mapping theorem, see [33]. This is also at the basis of the proof of the Donsker Theorem 13.

3.8.6 Proof of limiting shape results

In this section we prove the nonparametric BvM Theorem 14 and, as a fairly direct consequence given the results of [33], the Bayesian Donsker Theorem 13.

Proof of Theorem 14. The proof is similar to the corresponding proofs for Pólya trees or spike–and–slab Pólya trees, so we highlight only the few differences. The proof consists in two steps. First, proving convergence of finite–dimensional distributions and second, showing tightness of the rescaled posterior in a slightly smaller space.

Regarding convergence of finite–dimensional distributions, it suffices to note that for a fixed depth $L > 0$, the prior on wavelet coefficients of levels $l \leq L$ (for large enough n so that $\ell_0(n) > L$) coincides with the prior induced by a standard Pólya tree, for which the convergence of finite–dimensional distributions is shown in [29].

Regarding tightness, let $\bar{w} = (\bar{w}_l)$ be the sequence $\bar{w}_l = w_l / l^{\delta/2} = l^{2+\delta/2}$. This sequence is increasing in l and verifies $\bar{w}_l \gtrsim \sqrt{l}$, $\bar{w}_l = o(w_l)$ as $l \to \infty$, and $\bar{w}_{\ell_0(n)} \geq \log n$, using the assumption on $\ell_0(n)$. Now by the same argument as in the proof of Theorem 3 in [31], to establish the nonparametric BvM it suffices to prove that the distribution $\mathcal{L}(\sqrt{n}(f - C_n) \,|\, X)$ is tight in $\mathcal{M}_0(\bar{w})$, which is true if both laws $\mathcal{L}(\sqrt{n}(f - f_0) \,|\, X)$ and $\mathcal{L}(\sqrt{n}(f_0 - C_n))$ are tight.

Focusing first on the tightness of $\mathcal{L}(\sqrt{n}(f - f_0) \,|\, X)$, we wish to show that for any $\eta \in (0, 1)$, one can find $M = M(\eta)$ large enough such that

$$E_{f_0} \Pi[\|f - f_0\|_{\mathcal{M}_0(\bar{w})} > M/\sqrt{n} \,|\, X] \leq \eta. \tag{3.43}$$

We split, for $g = f - f_0$,

$$\|g\|_{\mathcal{M}_0(\bar{w})} \leq \max_{l \leq \ell_0(n),k} |g_{lk}|/\bar{w}_l + \max_{l > \ell_0(n),k} |g_{lk}|/\bar{w}_l =: (I) + (II).$$

For the term (I), as noted above, since the prior has a flat initialisation up to level $\ell_0(n)$, the induced prior and posterior on the first layers $l \leq \ell_0(n)$ of wavelet coefficients coincide with the prior/posterior of a standard Pólya tree, for which the corresponding tightness is proved in [29] (proof of Theorem 3). For the term (II), it follows from the proof of Theorem 11 (noting that the proof goes through with a prior with flat initialisation) that for T_n as in that proof and given $l > \ell_0(n)$, for any $\mathcal{T} \in T_n$ and on the event $\mathcal{B}_M$,

$$\int \max_{k:\,(l,k)\in\mathcal{T}_{int}} |f_{lk} - f_{0,lk}| d\Pi(f \,|\, \mathcal{T}, X) \leq C\sqrt{\frac{\log n}{n}}$$

and

$$\max_{k:\,(l,k)\notin\mathcal{T}_{int}} |f_{0,lk}| \leq C\frac{\log n}{\sqrt{n}}.$$

Since $\bar{w}_{\ell_0(n)} \geq \log n$ as verified above, one deduces that for any $\mathcal{T} \in T_n$ and on $\mathcal{B}_M$ the term (II) above is $O(1/\sqrt{n})$. Putting pieces together what precedes implies, with $\mathcal{E} = \{f_{\mathcal{T}},\ \mathcal{T} \in T_n\}$ as in the proof of Theorem 11,

$$\int_{\mathcal{E}} \|f - f_0\|_{\mathcal{M}_0(\bar{w})} d\Pi(f \,|\, X) = O_{P_0}(1/\sqrt{n}),$$

which in turn implies (3.43) using $\Pi[\mathcal{E}^c \,|\, X] = o_{P_0}(1)$.

It remains to prove tightness of $\mathcal{L}(\sqrt{n}(f_0 - C_n))$ in $\mathcal{M}_0(\bar{w})$. Again, one splits along indices: for $l \leq \ell_0(n)$, the posterior median tree estimator has same wavelet coefficients as the empirical measure P_n, and the estimate

$$E_{P_0} \max_{l \leq \ell_0(n)} \max_k |\langle P_0 - P_n, \psi_{lk}\rangle|/\bar{w}_l \leq C/\sqrt{n}$$

follows from the proof of Theorem 1 in [33] (see equation (36) there and lines below). For $l > \ell_0(n)$, one invokes the properties of the median tree estimator, namely

$$\max_{l > \ell_0(n)} \max_k |\hat{f}_{\mathcal{T}^*,lk} - f_{0,lk}| = O_{P_0}\left(\frac{\log n}{\sqrt{n}}\right), \tag{3.44}$$

as in Lemma 26, noting that the argument in that proof is unchanged for a prior with flat initialisation. This gives, using again $\bar{w}_{\ell_0(n)} \geq \log n$, that

$$\max_{l \leq \ell_0(n)} \max_k |\hat{f}_{\mathcal{T}^*,lk} - f_{0,lk}| = O_{P_0}(1/\sqrt{n}),$$

which gives the desired tightness property and concludes the proof. □

Remark 2. *It follows from the proof of Theorem 14 that there is quite some flexibility in the choice of the centering C_n. For instance, the projection $P_n(L_n)$ of the empirical measure P_n onto the first L_n levels of wavelet coefficients, with L_n the oracle supremum–norm cut–off $(n/\log n)^{1/(2\alpha+1)}$ can be used. This is because for $l \leq \ell_0(n)$ the projection $P_n(L_n)$ has by definition same wavelet coefficients as the empirical measure P_n, while for $l > \ell_0(n)$ equation (3.44) holds for $\langle P_n(L_n), \psi_{lk}\rangle$ instead of $f_{\mathcal{T}^*,lk}$ (with the even better bound $O_{P_0}(\sqrt{\log n/n})$), as in the proof of Theorem 1 in [33].*

Proof of Theorem 13. The results follows by applying Theorem 4 in [33]: since the posterior distribution on f satisfies the nonparametric BvM theorem 14, it suffices to check that the sequence (w_l) satisfies the condition $\sum_l w_l 2^{-l/2} < \infty$, which clearly holds, and to note that the centering $C_n = f_{\mathcal{T}^*}$ belongs to L^2. This shows that the Bayesian Donsker holds with centering $\hat{F}_n^{med} = \int_0^{\cdot} f_{\mathcal{T}^*}$. By using remark 2, the same result also holds with $\hat{F}_n^{med}$ replaced by the primitive, say $\mathbb{Z}_n(\cdot)$, of $P_n(L_n)$. But as noted in the proof of Corollary 1 in [33] (see also Remark 9 in [66]), we have $\|\mathbb{Z}_n - F_n\|_\infty = o_{P_0}(1/\sqrt{n})$, which implies the result with centering at F_n. □

3.8.7 Miscellaneous

We quickly remind that

$$\bar{Y}_\epsilon = E[Y_\epsilon \,|\, X^{(n)}] = \frac{a + N_X(I_{\epsilon 0})}{2a + N_X(I_\epsilon)}$$

and we define L_n as in (3.21).

Lemma 28. *Let $\alpha > 0$, $K > 0$ and P_0 be a distribution with a bounded density $f_0 \in \Sigma(\alpha, K)$ w.r.t. Lebesgue density. Then, for any*

$$M > \frac{1}{3}\left(\sqrt{\log 2}\sqrt{18\,\|f_0\|_\infty + \log 2} + \log 2\right),$$

the event

$$\mathcal{B}_M := \Big\{\forall l \geq 0, \quad \forall 0 \leq k \leq 2^l - 1, \\ M^{-1}|N_X(I_{l,k}) - nP_0(I_{l,k})| \leq \sqrt{\frac{n(l+L_n)}{2^l}} \vee (l + L_n) =: M_{n,l}\Big\}$$

is asymptotically certain under the law P_0 of the observations, i.e.

$$P_0\left(\mathcal{B}_M^c\right) = o(1).$$

Proof. According to Bernstein's inequality, for any $l \geq 0,\ 0 \leq k \leq 2^l - 1$,

$$P_0\left(|N_X(I_{l,k}) - nP_0(I_{l,k})| > MM_{n,l}\right) \leq 2\exp\left(-\frac{M^2 M_{n,l}^2/2}{nP_0(I_{l,k})(1 - P_0(I_{l,k})) + MM_{n,l}/3}\right).$$

By assumption, $P_0(I_{l,k})(1 - P_0(I_{l,k})) \leq \|f_0\|_\infty 2^{-l}$. Then, whenever $M_{n,l} = l + L_n$ (which is equivalent to $l + L_n \geq n2^{-l}$) or $M_{n,l} = \sqrt{\frac{n(l+L_n)}{2^l}}$, we can further upper bound the above quantity as

$$P_0\left(|N_X(I_{l,k}) - nP_0(I_{l,k})| > MM_{n,l}\right) \leq 2\exp\left(-\frac{M^2}{2\,\|f_0\|_\infty + 2M/3}(l + L_n)\right).$$

Therefore,

$$P_0\left(\mathcal{B}_M^c\right) \leq 2\sum_{l\geq 0} 2^l \exp\left(-\frac{M^2}{2\,\|f_0\|_\infty + 2M/3}(l + L_n)\right) = O(2^{-L_n}) = O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha+1}}\right),$$

the latter equality being true whenever

$$\frac{M^2}{2\,\|f_0\|_\infty + 2M/3} > \log 2,$$

i.e. $M > \frac{1}{3}\left(\sqrt{\log 2}\sqrt{18\,\|f_0\|_\infty + \log 2} + \log 2\right)$. □

Lemma 29. *Suppose $f_0 \in \Sigma(K, \alpha)$, with $0 < \alpha \leq 1$. For $M' > 0$, on the event $\mathcal{B}_M$ from Lemma 28, the set*

$$\mathcal{A} = \bigcap_{\epsilon : |\epsilon| < L_n} \left\{ |\bar{Y}_{\epsilon 0} - Y_{\epsilon 0}| \leq M' \sqrt{\frac{L_n}{n P_0(I_{\epsilon 0})}} \right\}$$

is such that

$$\Pi[\mathcal{A}^c \,|\, X] \lesssim \sum_{l \leq L_n} 2^l e^{-M'^2 \log n/4}$$

Proof. This proof comes from Lemmas 4 and 5 of [31]. For completeness, we give here some details of the proof. We have already explained that

$$Y_{\epsilon 0} \sim \text{Beta}(a + N_X(I_{\epsilon 0}), a + N_X(I_{\epsilon 0})).$$

We also noticed that on the event $\mathcal{B}_M$, $N_X(I_\epsilon) \to \infty$ uniformly for all $|\epsilon| \leq L_n$ for $n \to \infty$. Therefore, for n sufficiently large, $a + N_X(I_{\epsilon 0}) \wedge a + N_X(I_{\epsilon 0}) \geq 8$ for $|\epsilon| < L_n$. Also, under our assumptions, Lemma [2] from [29] allows us to say that, for n large enough, there exist μ, ν such that

$$0 < \mu \leq \frac{a + N_X(I_{\epsilon 0})}{2a + N_X(I_{\epsilon 0}) + N_X(I_{\epsilon 1})} \leq \nu < 1$$

uniformly on all $|\epsilon| < L_n$. In addition, if $i = |\epsilon|$, we have that

$$2a + N_X(I_{\epsilon 0}) + N_X(I_{\epsilon 1}) \geq N_X(I_{\epsilon 0}) \geq n P_0(I_{\epsilon 0}) - M\sqrt{2 n L_n 2^{-i}}.$$

Under our assumptions on f_0 and L_n, the last bound is itself lower bounded by $n P_0(I_{\epsilon 0})/2$ for n large enough. As a consequence, an application of Lemma 6 from [29] gives, for $x = M' L_n^{1/2}/2$,

$$\Pi\left[|\bar{Y}_{\epsilon 0} - Y_{\epsilon 0}| > \frac{x}{\sqrt{n P_0(I_{\epsilon 0})}} \,\middle|\, X \right] \leq D e^{-x^2/4}$$

for some constant D. Finally, a union bound helps us to conclude that

$$\Pi[\mathcal{A} \,|\, X] \lesssim \sum_{l \leq L_n} 2^l e^{-M'^2 \log n/4}.$$

□

Lemma 30 (Theorem 1.5 of [7])**.** *For any $x > 0$,*

$$a \left(\frac{x + 1/2}{e} \right)^{x+1/2} \leq \Gamma(x+1) \leq b \left(\frac{x + 1/2}{e} \right)^{x+1/2},$$

where Γ is usual Gamma function, and $a = \sqrt{2e}$ and $b = \sqrt{2\pi}$ are the best possible constants.

CHAPTER 4

Adaptive Wasserstein confidence sets

In the density estimation model, we investigate the problem of constructing *adaptive honest confidence sets* with diameter measured in Wasserstein distance W_p, $p \geq 1$, and for densities with unknown regularity measured on a Besov scale. As sampling domains, we focus on the $d-$dimensional torus $\mathbb{T}^d$, in which case $1 \leq p \leq 2$, and $\mathbb{R}^d$, for which $p = 1$. We identify necessary and sufficient conditions for the existence of adaptive confidence sets with diameters of the order of the regularity-dependent W_p-minimax estimation rate. Interestingly, it appears that the possibility of such adaptation of the diameter depends on the dimension of the underlying space. In low dimensions, $d \leq 4$, adaptation to any regularity is possible. In higher dimensions, adaptation is possible if and only if the underlying regularities belong to some interval of width at least $d/(d-4)$. This contrasts with the usual L_p-theory where, independently of the dimension, adaptation occurs only if regularities lie in a small fixed-width window. For configurations allowing these adaptive sets to exist, we explicitly construct confidence regions via the method of risk estimation. These are the first results in a statistical approach to adaptive uncertainty quantification with Wasserstein distances. Our analysis and methods extend to weak losses such as Sobolev norms with negative smoothness indices.

Table of Contents

4.1 Introduction

The construction of confidence sets is one of the fundamental problems of statistical inference, along with parameter estimation and hypothesis testing. Consider a model $\{P_f : f \in \mathcal{F}\}$, indexed by a family of functions $\mathcal{F}$, and observe (some quantity n of) data from the true distribution P_{f_0}, where $f_0 \in \mathcal{F}$. For most applications, having a single point estimate $\hat{f}_n$ of the true parameter f_0 is not enough, and one desires to evaluate its performance in terms of a loss function, that is, to know how far it lies from f_0. Producing a random set $C_n \subset \mathcal{F}$ from the data containing f_0 with a prescribed high probability $1-\alpha$ achieves this aim. In this work, we investigate the existence of *adaptive honest confidence sets*. Since f_0 is unknown, we must insist that C_n possesses the previous property not just for f_0, but for all $f \in \mathcal{F}$: we say that the confidence set C_n is *honest* if, at least for all sufficiently large n,

$$\inf_{f \in \mathcal{F}} P_f(f \in C_n) \geq 1 - \alpha.$$

Furthermore, we desire the diameter of the set C_n to shrink in n as quickly as possible; however, typically the precise speed of this shrinkage depends on aspects of the unknown density f_0 such as its regularity, and so we find ourselves in an adaptation problem.

We work in a *density estimation model*: consider observations $X_1, \ldots, X_n$ independent and identically distributed (i.i.d.) from a probability measure P_{f_0} with probability density f_0. The sample space of the X_i's will either be the $d-$dimensional torus $\mathbb{T}^d$ or $\mathbb{R}^d$. We then study procedures in a representative 'two-class adaptation problem', where f_0 belongs to one of two classes $\mathcal{F}(r)$ and $\mathcal{F}(s)$ (to be precisely defined below), indexed by regularity parameters $r < s$, such that $\mathcal{F}(s) \subset \mathcal{F}(r)$. An adaptive honest confidence set C_n should satisfy the above honest coverage condition, and also have a diameter that shrinks at the minimax estimation rate of whichever class f_0 belongs to (typically the rate is faster for the smaller class $\mathcal{F}(s)$). The construction of such a confidence set involves assessing the accuracy with which one can estimate f_0, which turns out to be more challenging than point estimation, as qualitative aspects of the parameter need to be identified. This problem has primarily been studied for L_p or related distances [23, 24, 25, 77, 85, 112, 140]. In L_2 loss, adaptive honest confidence sets exist only if the regularity parameters of interest lie in some 'small' interval ([23, 24, 85, 140]). More troublesome is the case of pointwise or L_∞ loss, where no such procedures exist ([77, 112]). This starkly contrasts the situation of adaptive estimation, where (perhaps at the cost of a logarithmic factor) it is possible to construct estimators which adapt to any regularity parameter ([53, 103]). Informally, these negative results come from the fact that, in L_2 loss, a related testing problem is easier (admits a faster convergence rate) than estimation, whereas for L_∞ loss, the testing and estimation problems are equally difficult ([23, 77]). This distinction highlights how the existence of adaptive honest confidence sets

depends on the geometry induced by the loss function (see [69, Chapter 8] for an overview of these results).

Arising from the ideas of Optimal Transport [87, 120], Wasserstein distances W_p, $p \geq 1$, between probability measures have recently been studied in a wide array of fields such as optimization, machine learning, and statistics. For $p \geq 1$, the $p-$Wasserstein distance between μ and ν, probability measures on a metric space $(\mathcal{X}, d)$, is defined as

$$W_p(\nu, \mu) := \inf_{\pi \in \Pi(\nu,\mu)} \left(\int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\pi(x,y) \right)^{1/p},$$

with the infimum ranging over the set $\Pi(\nu, \mu)$ of measures on $\mathcal{X} \times \mathcal{X}$ with given marginals ν and μ. It quantifies the minimal cost, as measured by the metric d, to morph the distribution μ into ν. For measures P_f and P_g dominated by a common measure and with densities f and g, this also entails a distance between those densities, with $W_p(f, g) := W_p(P_f, P_g)$.

Not only do these distances possess desirable theoretical properties ([166]), as they take into account the geometry of the underlying sample space, but recent numerical developments ([131]) have led to increased use in practical applications. They therefore now play a prominent role in statistics (see, for example, the review [129]). The convergence of the empirical distribution in W_p-distance is a well-studied problem (it stretches back to [54], with definitive results on limit theorems for the $\mathbb{R}$ sample space in [47]; for state-of-the-art results, see [60, 170]). In dimensions $d \geq 3$, the convergence rate of the empirical distribution (without further structural assumptions) is $n^{-1/d}$, demonstrating that convergence in W_p suffers from the curse of dimensionality. When measures have densities, a result in [171] states that, for certain classes of densities, W_p compares with Besov norms of smoothness -1, a classical result for the W_1 distance due to the Kantorovich-Rubinstein duality formula. The convergence rates they obtain for regular densities using this comparison result, which lie closer to the parametric rate $n^{-1/2}$, highlight the importance of regularity of the signal in high-dimensional settings: to some extent, the curse of dimensionality can be mitigated by smoothness.

In addition, these rates are faster than the standard s-smooth nonparametric convergence rate $n^{-\frac{s}{2s+d}}$ for L_p loss, $1 \leq p < \infty$, reflecting the fact that Wasserstein distances are weaker than L_p distances. In this chapter, we obtain similar quantitative improvements for testing separation rates of nonparametric statistical hypotheses. From this, on the bounded sample space $\mathbb{T}^d$ we deduce new *qualitative phenomena regarding the existence and non-existence of adaptive honest confidence sets* when using the loss functions W_p, $1 \leq p \leq 2$. Surprisingly, in dimensions $d \leq 4$ we construct confidence sets that can adapt to *any* set of regularities. This contrasts significantly with the fundamental limitations of adaptive confidence sets in L_p. In higher dimensions $d > 4$, adaptation is still possible for regularities belonging to a certain interval, which is wider than in the L_p case. The reason for this phenomenon is that while both the testing and estimation rates are faster than for L_p, the testing rate accelerates more, leaving 'more space' for adaptation to occur than in the analogous problem for L_p loss. As for densities on an unbounded sample space such as $\mathbb{R}^d$, the same phenomenon occurs, though we currently only have results for the W_1 distance.

From the general theory of confidence sets, it is known that such impossibility results may be circumvented if one is willing to remove certain 'troublesome' portions of the parameter space ([69, Proposition 8.3.7]). Recent works have focused on describing maximal sets of densities for which adaptation is possible, by introducing further structural constraints on the model. In [23], an L_2-adaptative confidence set is shown to exist if one discards all densities

within some positive radius of the smoother subclass; this radius converges to 0 as $n \to \infty$. Self-similarity conditions, which roughly speaking describe functions that are as regular at small scales as at larger scales, have been employed in the regression setting in [126, 132], as well as in density estimation [22, 36, 67]; see also [4]. In the study of adaptive Bayesian credible sets, self-similarity conditions were deployed in [139], and slightly more general polished tail assumptions were used in [143, 153]. We refer the interested reader to the review [117] for a more complete picture of the literature on uncertainty quantification in density estimation. In this chapter, we do not pursue such a programme for the Wasserstein distances, instead leaving this problem for future work.

The chapter is organized as follows. Section 4.2 formalizes our problem on the potential existence of adaptive honest confidence sets, and states our main results. The construction of such sets, whenever possible, and non-existence results are presented in Section 4.3 for the bounded sample space $\mathbb{T}^d$ and Section 4.4 for the unbounded sample space $\mathbb{R}^d$. Proofs are deferred to Sections 4.6 and 4.7.

4.2 Main Results

4.2.1 Setting and Definitions

Initially, we assume that f_0 is a density on the d-dimensional torus, $\mathbb{T}^d$, which may be identified with $(0,1]^d$. Our results also apply to the case of the unit cube $[0,1]^d$ (and hence any bounded rectangular subset of $\mathbb{R}^d$), which is the focus of [171]; see Section 4.5.3 below. For our loss function, we take the distance W_2; as described in Remark 4, this distance dominates W_p for $1 \leq p < 2$, in particular the important case of W_1. Later, we consider the situation where f_0 is a density on the whole of $\mathbb{R}^d$; while a study for $W_p, p > 1$ is beyond the scope of the present work, we obtain some definitive results for the loss function W_1 in Section 4.4.

Parameter Spaces

Here we define the classes of probability densities on $\mathbb{T}^d$ we consider; definitions for $\mathbb{R}^d$ are similar but deferred to Section 4.4. Let $\left\{\phi \equiv 1, \psi_{lk} : l \geq 0, 0 \leq k < 2^{ld}\right\}$ be an S-regular periodised Daubechies wavelet basis of $L_2(\mathbb{T}^d)$; see Appendix 4.5 for further details. We denote by $\langle f, g\rangle = \int_{\mathbb{T}^d} fg$ the usual inner product on L_2. For any $f \in L_p(\mathbb{T}^d), 1 \leq p < \infty$, the wavelet expansion

$$f = \langle f, 1\rangle + \sum_{l \geq 0} \sum_{k=0}^{2^{ld}-1} \langle f, \psi_{lk}\rangle \psi_{lk} \tag{4.1}$$

converges in L_p, and if f is continuous then the expansion converges uniformly on $\mathbb{T}^d$. We write $K_j(f)$ for the projection of f onto the first j resolution levels, i.e.

$$K_j(f) = \langle f, 1\rangle + \sum_{l < j} \sum_{k=0}^{2^{ld}-1} \langle f, \psi_{lk}\rangle \psi_{lk}. \tag{4.2}$$

To define the parameter classes, we use the scale of *Besov spaces*, $B^s_{pq}, 1 \leq p, q \leq \infty, s \geq 0$ as defined in Appendix 4.5. The index s should be interpreted as a smoothness or regularity parameter. Using the definition of the Besov norm (4.32) and the embedding $\ell_q \subset \ell_\infty$, for $f \in B^s_{pq}(\mathbb{T}^d)$ we have that

$$\|\langle f, \psi_{l\cdot}\rangle\|_p \leq \|f\|_{B^s_{pq}} 2^{-l\left(s + \frac{d}{2} - \frac{d}{p}\right)}. \tag{4.3}$$

Thus $f \in B^s_{pq}$ if its wavelet coefficients decay sufficiently fast as l grows, as measured by s.

The use of subsets of Besov spaces as parameter spaces in nonparametric statistics is well-established, and the scale contains several of the regularity classes usually considered in such settings: for example, the Sobolev spaces ($H^s = B^s_{22}$) and the Hölder spaces (for $s \notin \mathbb{N}, C^s = B^s_{\infty\infty}$, and for $s \in \mathbb{N}, C^s \subsetneq B^s_{\infty\infty}$). See [69, Section 4.3] for further discussion on this subject.

In standard loss functions such as L_2 or L_∞, it is typically assumed that f lies in some norm-ball in B^s_{pq}, for some choice of s, p, q. Here we slightly restrict the function class, insisting that the densities under consideration are bounded and bounded away from 0. In particular, the lower bound condition facilitates the faster minimax estimation rates of Proposition 6; it is shown in [171] that removing this condition results in slower rates for most parameter configurations.

Definition 7. *Let $1 \le p, q \le \infty$, $s \ge 0$, $B \ge 1$, $M \ge 1 \ge m > 0$. Define the function class*

$$\mathcal{F}_{s,p,q}(B; m, M) = \left\{ f \in B^s_{pq} : \int_{\mathbb{T}^d} f = 1, \quad \|f\|_{B^s_{pq}} \le B, \quad m \le f \le M \text{ a.e.} \right\}; \tag{4.4}$$

Note that we always have $1 \in \mathcal{F}_{s,p,q}(B; m, M)$, and so the class is non-empty. Henceforth we fix $p = 2$ and consider q, B, m, M to be given. Define

$$\mathcal{F}(s) := \mathcal{F}_{s,2,q}(B; m, M).$$

For large s and smaller values of $B \ge 1$, the condition $f \le M$ is superfluous. However, the imposition of the uniform lower bound $f \ge m > 0$ means that $\mathcal{F}(s)$ is a strict subset of the more typical parameter space $\{f \in B^s_{2q} : f \ge 0, \int f = 1, \|f\|_{B^s_{2q}} \le B\}$. Also, it is clear from the definition (4.32) that the continuous embedding $B^s_{pq} \subset B^r_{pq}$ holds with operator norm 1, so $\mathcal{F}(s) \subset \mathcal{F}(r)$ for $r \le s$.

Notation

For a probability density f, let P_f and E_f denote respectively the probability and expectation when $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f$. For real numbers a, b, we write $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. Given sequences (a_n) and (b_n), we write $a_n \lesssim b_n$ if there exists a constant $C > 0$ that is independent of n such that for all n, $a_n \le C b_n$; we also write $a_n \simeq b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Given any subset A of a metric space $(\mathcal{A}, d)$, we write $|A|_d$ for the d-diameter of A, defined by

$$|A|_d := \sup_{x,y \in A} d(x, y).$$

Given a subset $B \subset \mathcal{A}$ and a point $a \in \mathcal{A}$, we define the distance of a to B as

$$d(a, B) := \inf_{b \in B} d(a, b).$$

4.2.2 Description of the Problem

Suppose initially that $f \in \mathcal{F}(r)$ for some given $r \ge 0$. We wish to construct a confidence set C_n for the unknown density f; informally, we would like C_n to contain f with (some chosen) high probability. Specifically, given $\alpha \in (0, 1)$, we require any confidence set $C_n =$

$C_n(\alpha, X_1, \dots, X_n)$ to have *honest coverage* at level $1-\alpha$ over the class $\mathcal{F}(s)$, that is, there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$,

$$\inf_{f \in \mathcal{F}(r)} P_f(f \in C_n) \geq 1 - \alpha. \tag{4.5}$$

The 'honesty' refers to the uniformity over $\mathcal{F}(r)$. We remark that in the minimax paradigm, one must necessarily insist on honesty, since the true density f_0 is unknown: 'dishonest' adaptive confidence sets exist (see [69, Corollary 8.3.10]), but the index n_0 from which coverage is valid depends on the unknown f, so such procedures produce questionable guarantees in practice.

It is clear that the smaller the set C_n, the more informative it is; otherwise one could just take C_n to be the whole parameter space $\mathcal{F}(r)$. Thus we desire the W_2-diameter of our set C_n to shrink as quickly as possible in n. Suppose C_n satisfies the honest coverage condition (4.5) for some $\alpha \in (0,1)$, and let r_n be a positive sequence such that for some $\beta > \alpha$ and every $n \geq n_0$, we have

$$\inf_{\tilde{f}_n} \sup_{f \in \mathcal{F}(r)} P_f\Big(W_2(\tilde{f}_n, f) \geq r_n\Big) \geq \beta. \tag{4.6}$$

Here, the infimum is taken over all *estimators* (i.e. measurable functions) $\tilde{f}_n = \tilde{f}_n(X_1, \dots, X_n)$. Then by Lemma 2 in [140], the W_2-diameter of C_n satisfies, for $n \geq n_0$,

$$\sup_{f \in \mathcal{F}(r)} P_f\left(|C_n|_{W_2} \geq r_n\right) \geq \beta - \alpha;$$

in particular, its diameter cannot shrink faster than r_n with high probability. We define the *minimax estimation rate* (in probability) over $\mathcal{F}(s)$, denoted $r_n^*(s)$, to be the 'slowest' sequence (i.e. the largest such sequence up to a multiplicative prefactor) r_n such that (4.6) is satisfied for some $\beta > 0$ and some $n_0 \geq 1$. Usually this rate depends on the smoothness parameter s.

Remark 3. *The term 'minimax estimation rate' is often reserved for any sequence $\bar{r}_n$ such that*

$$\inf_{\tilde{f}_n} \sup_{f \in \mathcal{F}(r)} E_f W_2(\tilde{f}_n, f) \simeq \bar{r}_n.$$

By Markov's inequality, we have that $r_n^* \lesssim \bar{r}_n$. In fact, as shown by Proposition 6 below, in this problem the rates r_n^* and $\bar{r}_n$ coincide (possibly up to a logarithmic factor when $d = 2$).

In general, it is unrealistic to assume that the regularity r is known. Thus we find ourselves in an adaptation problem, where we wish to construct procedures that do not depend on the unknown smoothness r, but which result in (near-)optimal performance for a range of values of r. In order to highlight the main ideas, let us consider the two class adaptation problem, where for some fixed $s > r \geq 0$ we consider the model $\mathcal{F}(r)$, but also seek optimal performance over the smoother subclass $\mathcal{F}(s) \subset \mathcal{F}(r)$. We discuss after Theorem 18 how one might construct confidence sets adapting to a continuous window of smoothnesses $[r, R]$ or even all $r \geq 0$ simultaneously.

Definition 8. *We say that $C_n = C_n(\alpha, \alpha', X_1, \dots, X_n)$ is a* near-optimal adaptive W_2 confidence set over $\mathcal{F}(s) \cup \mathcal{F}(r)$, $s > r$, *if it satisfies the following properties, for given $\alpha, \alpha' \in (0,1)$:*

i) ***Honest Coverage:*** *for all n sufficiently large,*

$$\inf_{f \in \mathcal{F}(r)} P_f(f \in C_n) \geq 1 - \alpha; \tag{4.7}$$

ii) ***Diameter Shrinkage:*** *there exists a constant* $K = K(\alpha') > 0$ *such that*

$$\sup_{f \in \mathcal{F}(r)} P_f(|C_n|_{W_2} > KR_n(r)) \leq \alpha' \tag{4.8}$$

and

$$\sup_{f \in \mathcal{F}(s)} P_f(|C_n|_{W_2} > KR_n(s)) \leq \alpha', \tag{4.9}$$

for n *large enough, where the rate sequences* $R_n(r)$ *and* $R_n(s)$ *satisfy*

$$R_n(r) \leq a_n r_n^*(r) \quad \text{and} \quad R_n(s) \leq a_n r_n^*(s),$$

for a_n *some power of* $\log n$*, and* $r_n^*(r)$ *and* $r_n^*(s)$ *the minimax rates of estimation over* $\mathcal{F}(r)$ *and* $\mathcal{F}(s)$ *respectively (these are given in Proposition 6 for the case of* $\mathbb{T}^d$ *and Theorem 20 for the case of* $\mathbb{R}^d$*).*

Typically, for optimal adaptive confidence sets one insists that the rates $R_n(r), R_n(s)$ in (4.8) and (4.9) are equal up to constants to the minimax estimation rates $r_n^*(r), r_n^*(s)$. Our definition of 'near-optimal' allows for $R_n(t)$ to equal $r_n^*(t), t = r, s$, up to a logarithmic factor in n, and is thus a slight relaxation. Admitting this relaxation does not alter the (existence and) non-existence results of [23, 25, 69, 77], since these results are due to a polynomial discrepancy between minimax estimation and testing rates; see Section 4.3.3 below.

We only consider the problem of adaptation in the smoothness parameter and do not address the question of adaptation to other parameters in the definition of the class $\mathcal{F}(s)$, such as the Besov norm bound B. See Remark 7 below for a discussion of this issue.

4.2.3 Adaptive W_2 Confidence Sets on $\mathbb{T}^d$

Our first theorem exhaustively classifies the parameter configurations for which adaptive honest confidence sets exist for W_2 loss; in the cases where such confidence sets do exist, an explicit construction is given in Theorem 18 below.

Theorem 15. *Fix* $1 \leq q \leq \infty$, $B \geq 1$, $M \geq 1 \geq m > 0$. *Consider the two class adaptation problem for confidence sets as defined by (4.7)-(4.9).*

i) Let $d \leq 4$ *and* $s > r \geq 0$. *Then for any* $\alpha, \alpha' > 0$*, there exists a near-optimal adaptive* W_2 *confidence set.*

ii) Let $d > 4$ *and* $0 \leq r < s \leq \frac{2d-4}{d-4} r + \frac{d}{d-4}$. *Then for any* $\alpha, \alpha' > 0$*, there exists a near-optimal adaptive* W_2 *confidence set.*

iii) Let $d > 4$ *and* $0 \leq r < s$ *with* $s > \frac{2d-4}{d-4} r + \frac{d}{d-4}$. *Then for any* $\alpha, \alpha' > 0$ *such that* $2\alpha + \alpha' < 1$*, no near-optimal adaptive* W_2 *confidence set exists.*

Remark 4. *We have focussed on the particular choice of* W_2*; by Jensen's inequality, this distance dominates* W_p *for* $1 \leq p < 2$. *Since the minimax estimation rates in these problems are independent of* p *(c.f. Proposition 6), this means that the above existence results hold for* $W_p, 1 \leq p \leq 2$*, in particular for the important case of* W_1. *Moreover, in the case of* W_1*, one may remove the lower bound condition in the definition of* $\mathcal{F}(s)$*; see Remark 5 below.*

Theorem 15 says that in low dimensions, $d \leq 4$, there exists a confidence set which adapts optimally in W_2-diameter to *any* two smoothnesses $s > r \geq 0$. As the construction does not depend on s, in fact adaptation occurs simultaneously for all $s \geq r$ (strictly speaking, $r \leq s \leq S$ where S is the regularity of the wavelet basis used), where r is a chosen 'baseline' smoothness. Contrast this to the case of L_p loss, $2 \leq p \leq \infty$: for $p < \infty$, in any dimension, there exists a (near-)optimal adaptive confidence set if and only if $s \leq \frac{p}{p-1} r$ [23, 25]; for L_∞ loss, adaptive confidence sets do not exist for any choice of $s > r \geq 0$ [77, 112]. See [69, Section 8.3] for a complete account of the L_2 and L_∞ theory.

In higher dimensions $d > 4$, Theorem 15 (together with the same confidence set as constructed in the case $d \leq 4$) gives a continuous 'window' of smoothnesses for which adaptation occurs simultaneously, in a similar vein to the case of $L_p, p < \infty$. However, for the W_2 loss this window is significantly wider; moreover, regardless of how small we choose the minimal smoothness $r \geq 0$, this window has width at least $\frac{d}{d-4}$, whereas for $L_p, 2 \leq p < \infty$, the window is of width $\frac{r}{p-1} \leq r$, which will be very narrow for small values of r.

These results are related to the fact that W_2 is a weaker loss function than L_p: specifically, Proposition 5 and (4.12) show that on the class $\mathcal{F}(s)$, W_2 is comparable to a Sobolev (or Besov) norm of smoothness -1. In very low dimensions $d = 1, 2$, the estimation rate is independent of the smoothness parameter s, meaning that any confidence set satisfying (4.8) automatically satisfies the faster shrinkage condition (4.9) (with a possibly enlarged constant K). In low dimensions $d = 3, 4$, one finds a very fast minimax testing separation rate, which is at least as fast as the parametric rate of estimation $n^{-1/2}$ (this is implied by the above existence results and Lemma 32 below). Even in higher dimensions, there is a substantial acceleration in the testing separation rate as compared to L_2 loss. Meanwhile, although there is also some acceleration in the estimation rates, the effect is not so pronounced. This explains the wider window of adaptation seen in Theorem 15 for W_2 loss, as compared to L_p loss: the greater discrepancy between testing and estimation rates gives more room for adaptation to take place.

Theorem 15 is proved in Section 4.3; we outline the arguments now. For the existence result, we use the method of constructing confidence sets via risk estimation as in [24, 85, 140]; see [69, Section 6.4] for a concise summary of these ideas. These methods require the loss function under consideration to be a Hilbert space norm. Accordingly, we upper bound W_2 by a suitable Sobolev-type norm for which one can perform risk estimation with fast convergence rates; moreover, the estimation rates for this dominating norm differ from those for W_2 by only a logarithmic factor. In particular, the notions of near-optimal adaptive confidence sets for these two loss functions are equivalent. The non-existence result is obtained using a testing argument as in [23], [77] and others, together with a lower bound for the minimax separation rate in a related testing problem. Moreover, the precise characterisation of the separation rate identifies a certain small subset of $\mathcal{F}(r)$ consisting of 'problematic' densities which, once removed, permit the existence of confidence sets (with honesty relative to a smaller set of densities), as in the previous two references. We discuss the existence of these more general confidence sets after Theorem 19. These theoretical results and constructions extends more generally to the study of adaptive honest confidence sets with negative Sobolev norm distances, and we discuss them in Section 4.2.5. For $p > 2$, [25] develops a construction of adaptive L_p-confidence sets whose radii are selected via testing. Though an extension of these ideas to W_p-confidence sets should be possible, we do not pursue it here as the methodology greatly differs from the one used in the present chapter.

4.2.4 Adaptive W_1 Confidence Sets on $\mathbb{R}^d$

The case of densities on $\mathbb{R}^d$ is also of great interest; there are several situations in which it is unrealistic to assume compact support of the density f. Accordingly, let $X_1, \dots, X_n$ be an i.i.d. sample drawn from some unknown density f on $\mathbb{R}^d$. We take the Wasserstein-1 distance W_1 to be our loss function. We generalise our methods from the case of $\mathbb{T}^d$ to produce adaptive confidence sets for f which adapt over similar function classes $\mathcal{G}(s)$, defined in (4.21) below and involving a constant L which uniformly bounds the exponential moments of the densities in $\mathcal{G}(s)$. The discussion following Theorem 15 is relevant in this context as well: in particular, since the confidence sets constructed in cases (i) and (ii) do not depend on s, adaptation in fact takes place for the full range of possible values of s (i.e. $s \geq r$ when $d \leq 4$ and s in some given window when $d > 4$).

Theorem 16. *Fix $1 \leq q \leq \infty$, $B \geq 1$, $M \geq 1 \geq m > 0$. Consider the two class adaption problem for confidence sets defined by (4.7)-(4.9), with function classes $\mathcal{F}$ replaced by $\mathcal{G}$ and W_2 in place of W_1.*

i) Let $d \leq 4$ and $s > r \geq 0$. Then for any $\alpha, \alpha' > 0$, there exists a near-optimal adaptive W_1 confidence set.

ii) Let $d > 4$ and $0 \leq r < s \leq \frac{2d-4}{d-4} r + \frac{d}{d-4}$. Then for any $\alpha, \alpha' > 0$, there exists a near-optimal adaptive W_1 confidence set.

iii) Let $d > 4$, L be large enough and $0 \leq r < s$ with $s > \frac{2d-4}{d-4} r + \frac{d}{d-4}$. Then for any $\alpha, \alpha' > 0$ such that $2\alpha + \alpha' < 1$, no near-optimal adaptive W_1 confidence set exists.

The bound L on exponential moments in (4.21) is a technical condition which allows us to construct adaptive estimators and confidence sets via the method of risk minimization (see Section 4.4). We are naturally interested in the existence of confidence sets for large L, i.e. on larger classes of densities. Moreover, small values of L may lead to empty classes (see the discussion after Definition 10 below) for which the theory of confidence sets is superfluous.

4.2.5 Extension to negative Sobolev norm distances

To better understand the phenomena in Theorems 15 and 16, it is elucidating to consider negative order Sobolev norm loss, $H^{-t} = B_{22}^{-t}, t > 0$ (see Appendix 4.5 for definitions), since the W_2 distance is dominated by such a norm (see (4.12) below). One finds that the minimax estimation rate for $t \geq d/2$ is (up to a log factor) $n^{-1/2}$, so no meaningful adaptation is required and one constructs a confidence set which 'adapts' over all smoothnesses as in Proposition 7 below. When $t < d/2$, computations analogous to those in Section 4.3 show that the gap between testing and estimation rates are wider for larger t, enabling adaptation over a larger window of regularities (see Remark 10 below). Here, one finds a continuous transition as t increases from 0 (which is the L_2 case) to $d/2$, at which point confidence sets can adapt to any two smoothnesses. However, the specific geometry of the parameter space induced by the loss function is crucial, rather than how weak the loss function is *per se*: if instead we consider $B_{\infty\infty}^{-t}$ loss, when $t < d/2$ the minimax estimation and testing rates can be shown to coincide; meanwhile, the estimation rate is independent of the smoothness parameter when $t \geq d/2$. So in the case of $B_{\infty\infty}^{-t}$ loss, when $t < d/2$ no adaptive confidence sets exist for *any* two smoothnesses by Lemma 32 below, but for $t \geq d/2$ they trivially exist.

Whenever they exist, the construction of confidence sets in Section 4.3 below extends easily to the case of negative order Sobolev norms $H^{-t}, t > 0$, and other Besov norms using norm embeddings as in [69, Section 4.3]; see Remark 10 below.

4.3 Proof of Theorem 15

4.3.1 A Hilbert Norm Upper Bound for W_2

We wish to construct confidence sets by performing risk estimation. The inner product structure of Hilbert space norms makes them particularly amenable to risk estimation, and so we seek some Hilbert norm which upper bounds the W_2 distance.

For this, we introduce the *logarithmic Sobolev norm* ([69, Section 4.4]; see [32, 33] for another statistical application of such norms).

Definition 9. *Define the $H^{-1,\delta}$ norm of $f \in L_2(\mathbb{T}^d)$ as*

$$\|f\|_{H^{-1,\delta}} = |\langle f, 1\rangle| + \left(\sum_{l\geq 0} 2^{-2l} \max(l,1)^{2\delta} \|\langle f, \psi_{l\cdot}\rangle\|_2^2\right)^{1/2}.$$

Note the similarity to the definition of the $B_{22}^{-1} = H^{-1}$ norm given by (4.32); indeed, when $\delta = 0$ the two norms coincide with the Sobolev norm of regularity -1. We refer to this as a 'logarithmic' Sobolev space because the parameter δ measures the smoothness of f on a logarithmic scale.

We require the following comparison inequality from [171].

Proposition 5 (Theorem 3, [171])**.** *Let $1 \leq p < \infty$. Let f, g be two densities in $L_p(\mathbb{T}^d)$, and assume that for almost every $x \in \mathbb{T}^d$, $M \geq \max(f(x), g(x)) \geq m > 0$, for real numbers M and m. Then*

$$M^{-1/p'} \|f - g\|_{B_{p\infty}^{-1}} \lesssim W_p(f,g) \lesssim m^{-1/p'} \|f - g\|_{B_{p1}^{-1}}, \tag{4.10}$$

where $\frac{1}{p} + \frac{1}{p'} = 1$, and the constants depend only on d, p and the wavelet basis. Moreover, when $p = 1$, one may choose $m = 0$ (with the convention $0^0 = 1$).

This result is an extension of the celebrated Kantorovich-Rubinstein duality formula, which states that for two *probability measures* μ, ν on $\mathbb{T}^d$,

$$W_1(\mu,\nu) = \sup_{h \in \mathrm{Lip}_1(\mathbb{T}^d)} \int h \, \mathrm{d}(\mu - \nu), \tag{4.11}$$

where the supremum is taken over all functions $h : \mathbb{T}^d \to \mathbb{R}$ with Lipschitz constant bounded by 1. We may relate this to (4.10) using the sequence of norm-continuous embeddings ([69, Section 4.3])

$$B_{11}^{-1} \subset \left(B_{\infty\infty}^1\right)^* \subset BL(\mathbb{T}^d)^* \subset \left(B_{\infty 1}^1\right)^* \subset B_{1\infty}^{-1},$$

where $BL(\mathbb{T}^d)$ is the space of bounded Lipschitz functions on $\mathbb{T}^d$ (note that any Lipschitz function on $\mathbb{T}^d$ is bounded, so $BL(\mathbb{T}^d)$ and $\mathrm{Lip}_1(\mathbb{T}^d)$ coincide). However, in order to generalise this to $W_p, p > 1$, one must impose that the probability measures have densities which are bounded and bounded away from zero; indeed, for densities not bounded below, no norm

provides a similar comparison to W_p ([171, Theorem 7]), and convergence rates are slower than those in Proposition 6. Thus the restriction from the usual choices of Besov norm-balls to the classes $\mathcal{F}(s), s \geq 0$ is necessary.

A simple application of the Cauchy-Schwarz inequality confirms that $H^{-1,\delta} \subset B_{21}^{-1}$ as soon as $\delta > 1/2$. Thus in conjunction with the upper bound in Proposition 5, we have that, for $r \geq 0$, $f \in \mathcal{F}(r)$ and $\tilde{f}_n$ any estimator of f,

$$W_2(f, \tilde{f}_n) \lesssim \|f - \tilde{f}_n\|_{B_{21}^{-1}} \lesssim \|f - \tilde{f}_n\|_{H^{-1,\delta}}, \tag{4.12}$$

where the first constant depends on the parameters of the class $\mathcal{F}(r)$, but the second constant depends only on the wavelet basis and d.

Remark 5. *When using W_1 loss, one may consider the class $\mathcal{F}(s)$ with the choice $m = 0$, i.e. densities are not required to be bounded away from zero. Then the $H^{-1,\delta}$ norm still provides an upper bound for W_1 for densities in $\mathcal{F}(s)$ due to the upper bound in (4.10) and the sequence of continuous embeddings $H^{-1,\delta} \subset B_{21}^{-1} \subset B_{11}^{-1}$, where the second embedding follows from Jensen's inequality (with operator norm 1).*

For the remainder of this section, we work in $H^{-1,\delta}$ risk; as soon as $\delta > 1/2$, this provides a Hilbert norm upper bound for the W_2 risk. In particular, any coverage guarantee for a $H^{-1,\delta}$ ball is automatically inherited by the W_2 ball with the same centre and radius scaled by the embedding constant from (4.12). Of course, by constructing confidence sets for a stronger loss function, we may not be able to attain near-optimal diameter shrinkage, but we shall see that this is not the case.

4.3.2 Construction of Confidence Sets

We first give the minimax estimation rates for the problem under consideration. These are important for two reasons: firstly, they provide the benchmark for the 'size' of an optimal confidence set. Moreover, our confidence sets are centred at a suitable estimator of f, which must perform well for the resulting confidence set to also have good performance. In the density estimation problem, the estimation rates for W_2 loss are as follows:

Proposition 6. *Let $s \geq 0$ and let $r_n^*(s)$ denote the minimax rate of estimation over $\mathcal{F}(s)$. Then*

$$r_n^*(s) \lesssim \begin{cases} n^{-1/2}, & d = 1, \\ n^{-1/2} \log n, & d = 2, \\ n^{-\frac{s+1}{2s+d}}, & d \geq 3, \end{cases}$$

where the constant depends on the parameters of the class $\mathcal{F}(s)$ and the wavelet basis. Moreover, for any $s \geq 0$,

$$r_n^*(s) \gtrsim \begin{cases} n^{-1/2}, & d = 1, 2 \\ n^{-\frac{s+1}{2s+d}}, & d \geq 3, \end{cases}$$

where the infimum is over all estimators $\tilde{f}_n$ based on a sample of size n.

The upper bounds follow from Theorem 1 in [171] and Remark 3. The lower bounds are proved as in Theorem 6.3.9 in [69], where one ensures the existence of a suitable W_2-separated set using the lower bound in Proposition 5. See also Theorem 2 in [171].

We centre our confidence sets at an estimator $\hat{f}_n$ of f which has near-optimal convergence over the classes $\mathcal{F}(s)$ and $\mathcal{F}(r)$. The theory of adaptive estimation is relatively complete, and in the vast majority of cases it is possible to construct adaptive estimators which converge at the minimax estimation rate (perhaps up to a logarithmic factor) over a wide range of smoothnesses - we mention only the classical references [53] and [103].

The choice of Wasserstein loss adds a minor complication to the usual case of 'norm-type' loss functions. The Wasserstein distance $W_p(f, \tilde{f}_n)$ is only well-defined if $\tilde{f}_n$ is also a density, and thus we ought to insist that any estimator we define is indeed a density almost surely. To achieve this, given any wavelet-based estimator of the form

$$\tilde{f}_n = \tilde{f}_{-1} + \sum_{l \geq 0} \sum_{k=0}^{2^{ld}-1} \tilde{f}_{lk} \psi_{lk}$$

where $\tilde{f}_{lk}$ are the wavelet coefficients of the estimator, we insist that $\tilde{f}_{-1} = 1$. This ensures that $\int_{\mathbb{T}^d} \tilde{f}_n = 1$. The problem of non-negativity is more subtle. In [171], it was addressed by projecting $\tilde{f}_n$ onto the class of densities $\mathcal{F}(r)$ with respect to the B^{-1}_{p1} norm, where r is the smallest regularity to which we want to adapt. However, this projection step makes the estimator essentially intractable. Instead, we use the well-known L_∞ consistency of the adaptive estimators considered below (c.f. [53], for example) together with the fact that the densities in $\mathcal{F}(r)$ are uniformly bounded away from 0 to conclude that for sufficiently large n, with high probability $\tilde{f}_n$ is in fact a probability density. Whenever $\tilde{f}_n$ fails to be non-negative, we simply replace it with an arbitrary choice of density (e.g. uniform); as $n \to \infty$, this event occurs with vanishing probability.

Theorem 17. *Let $d \geq 2$. Then there exists an estimator $\hat{f}_n$ of f such that for all $n \geq n_0(B)$ and all $s \geq 0$,*

$$\sup_{f \in \mathcal{F}(s)} E_f \|f - \hat{f}_n\|^2_{H^{-1,\delta}} \lesssim (\log n)^{2\delta} \left(\frac{n}{\log n} \right)^{-\frac{2(s+1)}{2s+d}},$$

where the constant depends on B, d and the wavelet basis.

The definition of $\hat{f}_n$ and proof of Theorem 17 can be found in Appendix 4.6, and follows from the classical ideas of [53].

Next, we introduce a U-statistic to perform risk estimation. Recall that given any estimator $\tilde{f}_n$ of f such that $\langle \tilde{f}_n, 1 \rangle = 1$, the $H^{-1,\delta}$ loss can be expressed as

$$\|f - \tilde{f}_n\|^2_{H^{-1,\delta}} = \sum_{l \geq 0} 2^{-2l} (l \vee 1)^{2\delta} \sum_{k=0}^{2^{ld}-1} \langle f - \tilde{f}_n, \psi_{lk} \rangle^2.$$

To estimate this loss, we use the approach of sample splitting. Suppose we have a sample of size $2n$ which we divide into two subsamples: $\mathcal{S}^1 = (X_1, \ldots, X_n), \mathcal{S}^2 = (X_{n+1}, \ldots, X_{2n})$. Denote expectation with respect to sample i by $E^{(i)}$; we denote variances and probabilities accordingly. We compute our estimator $\tilde{f}_n = \tilde{f}_n(X_1, \ldots, X_n)$ based on $\mathcal{S}^1$ and, for $j \geq 0$, define the U-statistic based on the sample $\mathcal{S}^2$ as

$$U_{n,j}(\tilde{f}_n) = \frac{2}{n(n-1)} \sum_{i<i', i, i' \in \mathcal{S}^2} \sum_{l<j} 2^{-2l} (l \vee 1)^{2\delta} \sum_{k=0}^{2^{ld}-1} \left(\psi_{lk}(X_i) - \langle \psi_{lk}, \tilde{f}_n \rangle \right) \left(\psi_{lk}(X_{i'}) - \langle \psi_{lk}, \tilde{f}_n \rangle \right). \tag{4.13}$$

Since the sample is i.i.d., we see that

$$E_f^{(2)} U_{n,j}(\tilde{f}_n) = \sum_{l<j} 2^{-2l}(l \vee 1)^{2\delta} \sum_{k=0}^{2^{ld}-1} \langle \psi_{lk}, f - \tilde{f}_n \rangle^2 = \|K_j(f - \tilde{f}_n)\|_{H^{-1,\delta}}^2.$$

Thus $U_{n,j}(\tilde{f}_n)$ is an unbiased estimator of the j^{th} resolution level approximation of the loss $\|f - \tilde{f}_n\|_{H^{-1,\delta}}$. The key idea behind the U-statistic is that the removal of the diagonal in the outermost sum in (4.13) eliminates the highest variance terms. Thus by averaging over $O(n^2)$ terms with small variance, we expect the U-statistic to have very small variance (as in Theorem 6.4.6 of [69]). This is confirmed by the next lemma.

Lemma 31. *Assume $f \in L^\infty(\mathbb{T}^d)$ is a probability density, and $\tilde{f}_n$ is an estimator for f based on the subsample $\mathcal{S}^1$. Then*

$$\begin{aligned} \mathrm{Var}^{(2)}(U_{n,j}(\tilde{f}_n)) &\leq \frac{4\|f\|_\infty}{n} \left(\max_{l \geq -1} 4^{-l}(1 \vee l)^{2\delta} \right) \|K_j(f - \tilde{f}_n)\|_{H^{-1,\delta}}^2 \\ &\quad + \frac{2\|f\|_\infty^2}{n(n-1)} \sum_{l \leq j-1} 2^{l(d-4)}(l \vee 1)^{4\delta} \\ &=: \kappa_{n,j,\delta}^2(f). \end{aligned} \tag{4.14}$$

This result is analogous to Theorem 4.1 in [140]; for completeness, we give a proof in Appendix 4.6.

With the adaptive estimator $\hat{f}_n$ and the U-statistic $U_{n,j}(\hat{f}_n)$ in hand, we are now ready to give the construction of optimal confidence sets for the two-class adaptation problem.

We first note that for $d = 1, 2$, the minimax rates of estimation from Proposition 6 do not depend on the smoothness parameter s; in particular, the two diameter shrinkage conditions (4.8) and (4.9) become a single condition. Thus in these dimensions, defining an adaptive confidence set is very easy; indeed, there is no meaningful adaptation which needs to take place.

When $d = 1$, the empirical measure is a minimax optimal estimator of the sampling measure (see, for instance, [60] or [170]). When $d = 2$, we centre at the adaptive estimator from Theorem 17 in place of the empirical measure P_n, as P_n is no longer minimax optimal, and standard kernel or wavelet projection estimators require choices of tuning parameters depending on the smoothness parameter to attain optimal rates.

Proposition 7. *i) Let $d = 1$. Consider the two-class adaptation problem over $\mathcal{F}(s) \cup \mathcal{F}(r)$ where $s > r \geq 0, q \in [1, \infty], B \geq 1, M \geq 1 \geq m > 0$ are all fixed. Then given any $\alpha \in (0, 1)$, the confidence set based on a sample $X_1, \ldots, X_n$ defined by*

$$C_n = \left\{ g \in \mathcal{F}(r) : W_2(P_g, P_n) \leq D\alpha^{-1/2} n^{-1/2} \right\}$$

is an optimal adaptive W_2 confidence set, where $P_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ is the n-sample empirical measure and the constant D depends on B, m and the wavelet basis.

ii) Let $d = 2$. Consider the two-class adaptation problem over $\mathcal{F}(s) \cup \mathcal{F}(r)$ where $s > r \geq 0, q \in [1, \infty], B \geq 1, M \geq 1 \geq m > 0$ are all fixed. Then given any $\alpha \in (0, 1)$, the confidence set based on a sample $X_1, \ldots, X_n$ defined by

$$C_n = \left\{ g \in \mathcal{F}(r) : W_2(g, \hat{f}_n) \leq D\alpha^{-1/2} n^{-1/2} (\log n)^{2+\delta} \right\}$$

is a near-optimal adaptive W_2 confidence set, where $\hat{f}_n$ is the adaptive estimator from Theorem 17 and the constant D depends on B, m and the wavelet basis.

The diameter shrinkage conditions are met trivially, while honest coverage follows from Chebyshev's inequality in a standard fashion.

When $d \geq 3$, the minimax rates depend on the smoothness parameter and so the diameter shrinkage condition differs between $\mathcal{F}(r)$ and $\mathcal{F}(s)$, $r \neq s$. In particular, this precludes any confidence set C_n with deterministic radius, as used above. Instead, we centre at the adaptive estimator $\hat{f}_n$ from Theorem 17, and use the estimate of its loss provided by the U-statistic $U_{j,n}(\hat{f}_n)$ as defined in (4.13) to determine the radius. We write $U_j := U_{j,n}(\hat{f}_n)$ in the sequel.

Theorem 18. *Let $d \geq 3$. Fix $B \geq 1, M \geq 1 \geq m > 0, 1 \leq q \leq \infty$, and let $s > r \geq 0$. If $d > 4$, assume additionally that $s \leq \frac{2d-4}{d-4} r + \frac{d}{d-4}$. Fix $\alpha \in (0,1)$, and $\delta > 1/2$. Consider the confidence set based on a sample of size $2n$, $\mathcal{S}^1 \cup \mathcal{S}^2$ given by*

$$C_n = \left\{ g \in \mathcal{F}(r) : \|g - \hat{f}_n^T\|_{H^{-1,\delta}} \leq \sqrt{z_\alpha \kappa_{n,j_n,\delta}(g) + U_{j_n} + G(j_n)} \right\} \tag{4.15}$$

where $\hat{f}_n^T$ is computed on $\mathcal{S}^1$, U_{j_n} is computed on $\mathcal{S}^2$ and:

- $\kappa^2_{n,j,\delta}(g) := \frac{4\|g\|_\infty}{n} \|K_j(g - \hat{f}_n^T)\|^2_{H^{-1,\delta}} + \frac{2\|g\|^2_\infty}{n(n-1)} \sum_{l \leq j-1} 2^{l(d-4)} (l \vee 1)^{4\delta}$;
- j_n *is such that* $2^{j_n} \simeq \left(\frac{n}{\log n}\right)^{\frac{1}{2r+d/2}}$;
- $G(j_n) = j_n^{2\delta} 2^{-2j_n(r+1)} \log n$;
- $z_\alpha = (\alpha/2)^{-1/2}$.

Then for all $n \geq n_0(B)$, C_n satisfies (4.7), as well as (4.8) and (4.9) for a suitable constant $K > 0$ depending on r, s, α, α' and the parameters of the class $\mathcal{F}(r)$ with the rates

$$R_n(r) = (\log n)^{\delta + \frac{r+1}{2r+d}} n^{-\frac{r+1}{2r+d}}, \qquad R_n(s) = (\log n)^{\delta + \frac{s+1}{2s+d}} n^{-\frac{s+1}{2s+d}}.$$

In particular, C_n is a near-optimal adaptive W_2 confidence set over $\mathcal{F}(s) \cup \mathcal{F}(r)$.

Remark 6 (Adaptation over ranges of classes)**.** *Note that the construction of C_n is completely independent of s, and $\hat{f}_n$ adapts simultaneously over all $s \geq 0$. So when $d \leq 4$, C_n adapts simultaneously over all $s \geq r$, and when $d > 4$, C_n adapts simultaneously over the full window of admissible values of s.*

Remark 7 (Adaptation to other parameters)**.** *We note that the construction of the confidence set in Theorem 18 does not depend on B or m, and so in fact this particular confidence set is also adaptive over $B \geq 1$ and $m > 0$, in the sense that any dependence of the minimax rates $r_n^*(r), r_n^*(s)$ on B or m are eventually accounted for by the logarithmic term in $R_n(r), R_n(s)$. (Note however that the constants in our theoretical guarantees explode as $B \to \infty$ or $m \to 0$.) However, the construction of C_n does depend on M. See [23] for more discussion on the role of M.*

Remark 8 (Adapting to wider ranges of smoothnesses in high dimensions)**.** *In the $d > 4$ case, following the ideas in [23], one may still obtain adaptation over a window of the form $[0, R]$ for arbitrary $R > 0$ at the cost of removing certain troublesome portions of the classes $\mathcal{F}(r), r \in [0, R]$. In this restricted model, one can identify the smoothness of the unknown density within a window of the form $\left[r, \frac{2d-4}{d-4}r + \frac{d}{d-4}\right]$ using tests as in [23] or [126]. Once this window is identified, in particular the relevant value of r, one can use the associated confidence set as constructed in Theorem 18.*

Remark 9. *(Necessity of log-factors) One may ask whether it is possible to remove the log-factors in the shrinkage rates and construct a confidence set with $R_n(r) = r_n^*(r), R_n(s) = r_n^*(s)$. These log factors fundamentally arise from the use of the embedding $H^{-1,\delta} \hookrightarrow B_{21}^{-1}$ for $\delta > 1/2$. For confidence sets constructed via risk estimation we conjecture that this is a necessary step, as it is precisely the accelerated risk estimation for Hilbert space norms which enables the adaptivity of the confidence set. However, it is conceivable that another approach, such as the testing method of [25], could be used to construct W_2 confidence sets with sharp diameter shrinkage rates (although such an approach will not generalise beyond the two class problem).*

Remark 10 (Weak Sobolev norms $H^{-t}, t > 0$)**.** *Our methods extend to the use of negative order Sobolev norms $H^{-t} = B_{22}^{-t}, t > 0$ as loss functions in place of $H^{-1,\delta}$ (see Appendix 4.5 for definitions). The analysis of the estimator $\hat{f}_n$ is completely analogous, and one must suitably augment the U-statistic $U_{n,j}$ to estimate the H^{-t} loss. One finds that the resulting confidence set $\tilde{C}_n$ adapts to any two smoothnesses $0 \leq r < s < \infty$ when $t \geq d/4$; if instead $t < d/4$, adaptation is possible over a window of smoothnesses $0 \leq r < s \leq \frac{d}{d-4t}t + \frac{2d-4t}{d-4t}r$. Moreover, in this latter case, the arguments of Section 4.3.3 below can be augmented to show that if s does not lie in this window, then no such confidence set can exist.*

The proof of this theorem proceeds similarly to that of Proposition 2.1 in [140], and is given in Appendix 4.6.

The confidence sets constructed above prove statements (i) and (ii) of Theorem 15.

4.3.3 Testing rates and non-existence of Confidence Sets

We turn now to proving the impossibility result (iii) in Theorem 15.

The question of existence of adaptive confidence sets is closely related to a composite hypothesis testing problem. This connection was identified in the first works on adaptive confidence sets; for a complete decision-theoretic overview, see [69, Chapter 8]. For $\rho \geq 0$ and $s > r \geq 0$, define the separated function class

$$\tilde{\mathcal{F}}(r, \rho) := \{f \in \mathcal{F}(r) : W_2(f, \mathcal{F}(s)) \geq \rho\}$$

We may have $\rho = 0$, in which case $\tilde{\mathcal{F}}(r, 0) = \mathcal{F}(r)$. However, if $\rho > 0$ then $\tilde{\mathcal{F}}(r, \rho)$ is a strict subset of $\mathcal{F}(r)$, disjoint from $\mathcal{F}(s)$. The testing problem we consider is

$$H_0 : f \in \mathcal{F}(s) \quad \text{vs.} \quad H_1 : f \in \tilde{\mathcal{F}}(r, \rho). \tag{4.16}$$

As the usefulness of a test is naturally assessed by the sum of its Type I and Type II errors, the minimax rate of testing for the problem (4.16) is defined as any sequence $(\rho_n^*)_{n \geq 1}$ such that

- For any $\beta' > 0$, there exists a constant $L = L(\beta')$ and a measurable test $\Psi_n : \left(\mathbb{T}^d\right)^n \to \{0,1\}$ such that

$$\sup_{f\in\mathcal{F}(s)} \mathbb{E}_f[\Psi_n] + \sup_{f\in\tilde{\mathcal{F}}(r,L\rho_n^*)} \mathbb{E}_f[1-\Psi_n] \le \beta'. \tag{4.17}$$

- There exists some $\beta > 0$ such that for all $\rho_n = o(\rho_n^*)$,

$$\liminf_{n\to\infty} \inf_{\Psi_n} \left[\sup_{f\in\mathcal{F}(s)} \mathbb{E}_f[\Psi_n] + \sup_{f\in\tilde{\mathcal{F}}(r,\rho_n)} \mathbb{E}_f[1-\Psi_n] \right] \ge \beta, \tag{4.18}$$

  where the infimum ranges over the set of tests Ψ_n.

The following result characterises the role of the minimax testing rate ρ_n^* in the existence and non-existence of confidence sets. Essentially, it says ρ_n^* provides a 'speed limit' on how quickly the confidence set can shrink when f is in the smoother submodel $\mathcal{F}(s)$:

Lemma 32 (Proposition 8.3.6, [69]). *Let ρ_n^* be the minimax testing rate for (4.16), and $\tilde{r}_n(s), \tilde{r}_n(r)$ be two sequences such that $\tilde{r}_n(s) = o(\rho_n^*)$ and $\tilde{r}_n(s) = o(\tilde{r}_n(r))$. Let $\alpha, \alpha' > 0$. Then, for any $\rho_n = o(\rho_n^*)$ and $L > 0$, there does not exist any set $C_n(\alpha, X_1, \dots, X_n)$ satisfying*

- $\liminf_{n\to\infty} \inf_{f\in\mathcal{F}(s)\cup\tilde{\mathcal{F}}(r,\rho_n)} P_f(f \in C_n) \ge 1-\alpha$,
- $\limsup_{n\to\infty} \sup_{f\in\tilde{\mathcal{F}}(r,\rho_n)} P_f\left(|C_n|_{W_2} > L\tilde{r}_n(r)\right) \le \alpha'$,
- $\limsup_{n\to\infty} \sup_{f\in\mathcal{F}(s)} P_f\left(|C_n|_{W_2} > L\tilde{r}_n(s)\right) \le \alpha'$,

as long as α, α' are such that $0 < 2\alpha + \alpha' < \beta$, with β as in (4.18).

This non-existence phenomenon occurs because any C_n satisfying the conditions of the Lemma induces a test

$$\Psi_n = \mathbb{1}\{C_n \cap \tilde{\mathcal{F}}(r, \rho_n') \neq \varnothing\}$$

which is uniformly consistent for the separation rate ρ_n' in the sense of (4.17) whenever $\rho_n = o(\rho_n')$. If we were able to choose ρ_n' to be $o(\rho_n^*)$, this would contradict the definition of the minimax testing rate ρ_n^*; thus no such confidence set can exist. Note that the argument works for any rate $\tilde{r}_n(s) = o(\rho_n^*)$, not just the minimax rate of estimation; in particular, we can multiply the minimax estimation rate by a poly-logarithmic factor so long as there is a polynomial gap between the testing and estimation rates.

It remains to determine the minimax rate of testing for the problem (4.16); this is done in the following theorem.

Theorem 19. *Assume $s > r \ge 0$ and $d > 4$. Let ρ_n^* be the minimax rate of testing for the problem (4.16). Then there exists a constant $c > 0$ depending on the parameters of the class $\mathcal{F}(s)$ and the wavelet basis, and $n_0 = n_0(B, M)$ such that for all $n \ge n_0$,*

$$\rho_n^* \ge c n^{-\frac{r+1}{2r+d/2}}.$$

Also, (4.18) holds for any $\beta < 1$.

The proof of Theorem 19 is given in Appendix 4.6, and follows a multiple-testing lower bound. Assume now that $d > 4$ and $s > \frac{2d-4}{d-4}r + \frac{d}{d-4}$. Then the minimax rate of testing ρ_n^* is slower than the minimax estimation rate $r_n^*(s)$ by a polynomial factor; in light of Lemma 32, this means there is no near-optimal adaptive W_2 confidence set over $\mathcal{F}(s) \cup \mathcal{F}(r)$ for any practical choice of α, α' (for such a set to exist, we would require $2\alpha + \alpha' \geq 1$). This proves statement (iii) of Theorem 15. However, this does not rule out the existence of confidence sets satisfying weaker conditions than those in Definition 8, namely those listed in Lemma 32 for some $\rho_n \geq L\rho_n^*$, $L > 0$. Such sets actually exists in view of Proposition 8.3.7 of [69] and Theorem 17.

Moreover, the confidence set C_n constructed in Theorem 18 in conjunction with the argument used to prove Lemma 32 shows that the lower bound of Theorem 19 is sharp up to a poly-logarithmic factor.

4.4 Extension of the Theory to $\mathbb{R}^d$

Having provided a fairly complete resolution of the problem of adaptive W_2 confidence sets when the sample space is $\mathbb{T}^d$, we extend our results to the case of the unbounded sample space $\mathbb{R}^d$ with W_1 loss. The key tool is the Kantorovich-Rubinstein duality formula ([86])

$$W_1(f,g) = \sup_{h \in \mathrm{Lip}_1(\mathbb{R}^d)} \int_{\mathbb{R}^d} h(x)(f(x) - g(x))\, \mathrm{d}x, \tag{4.19}$$

where $\mathrm{Lip}_1(\mathbb{R}^d)$ is the set of 1-Lipschitz functions on $\mathbb{R}^d$.

Our techniques do not extend to the distances W_p, $p > 1$, due to the dependence on the lower bound m in Proposition 5 (which is the analogue of (4.19) for $p > 1$): any density on $\mathbb{R}^d$ must decay to 0 at infinity, so using this result yields suboptimal convergence rates, even under favourable tail conditions.

In this section, it is assumed that we observe $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f_0$ for some density f_0 on $\mathbb{R}^d$, and we wish to perform inference on f_0 using W_1 as the loss function.

4.4.1 Parameter Spaces

We use an S-regular tensor product wavelet basis of $L^2(\mathbb{R}^d)$ of the form

$$\left\{\phi_k, \psi_{lk} : k \in \mathbb{Z}^d, l \geq 0\right\}$$

as introduced in Appendix 4.5 (we index the ψ_{lk} using only k, l by a slight abuse of notation). We write $K_j(f)$ for the projection of f onto the first j resolution layers, as in (4.2). Besov norms on $\mathbb{R}^d$, also defined in Appendix 4.5, are defined analogously to those on $\mathbb{T}^d$, and the relation (4.3) holds.

Our goal is to construct an adaptive confidence set for the true density f_0 using risk estimation, where the adaptation occurs with respect to the smoothness parameter s. We shall consider functions in B_{2q}^s. Unlike our previous classes $\mathcal{F}(s)$ on $\mathbb{T}^d$, we need not assume that our densities are bounded away from zero, or something analogous such as sufficiently slow decay in the tails. However, in order to deal with the unboundedness of the sample space $\mathbb{R}^d$, we must impose a moment condition.

For $\alpha, \beta > 0$, define the α, β-exponential moment of a density f as

$$\mathcal{E}_{\alpha,\beta}(f) := \int_{\mathbb{R}^d} \exp\left(\beta \|x\|^\alpha\right) f(x)\, \mathrm{d}x = E_f\left(e^{\beta \|X\|^\alpha}\right). \tag{4.20}$$

Definition 10. *Let $1 \le p, q \le \infty, s \ge 0, B \ge 1, M > 0, \alpha, \beta > 0$ and $L \ge 1$. Define the function class*

$$\mathcal{G}_{s,p,q}(B, M; \alpha, \beta, L) = \left\{ f \in B^s_{pq}(\mathbb{R}^d) : \int_{\mathbb{R}^d} f = 1,\ \|f\|_{B^s_{pq}} \le B,\ 0 \le f \le M \text{ a.e.},\ \mathcal{E}_{\alpha,\beta}(f) \le L \right\}. \tag{4.21}$$

Henceforth, we fix $p = 2$ and consider $q, B, M, \alpha, \beta, L$ to be given. Define

$$\mathcal{G}(s) := \mathcal{G}_{s,2,q}(B, M; \alpha, \beta, L).$$

Observe that for M close to 0 and L close to 1, the class $\mathcal{G}(s)$ is empty. We therefore assume in the sequel that L is sufficiently large (depending on M, B) for $\mathcal{G}(s)$ to be non-empty.

The focus on $p = 2$ is quite natural in view of the material developed in the previous section, relying on risk estimation to compute the diameter of confidence sets. Combining the exponential moment condition and the bound on the B^s_{2q}-norm, we prove in Lemma 38 that densities in $\mathcal{G}(s)$ also have their B^s_{1q}-norm bounded by a constant depending on the class parameters.

4.4.2 Estimation Upper Bounds for W_1

As before, we should insist on our estimator $\tilde{f}_n$ being a density almost surely. Indeed, the fact that $\tilde{f}_n$ has total mass 1 is vital to the proof of Proposition 8 below. However, we note that there is no intrinsic requirement in (4.19) that f and g should be nonnegative, and so we will allow our estimators to take negative values. If a genuine density is required, one can just take the positive part of the estimator and renormalize.

The following proposition gives an upper bound on the W_1 distance which is convenient for wavelet estimators.

Proposition 8. *For any probability density f with a finite first moment and any estimator $\tilde{f}_n$ of f which has a finite first moment almost surely, we have that*

$$W_1(\tilde{f}_n, f) \lesssim \sum_{k \in \mathbb{Z}^d} \|k\| |\langle f - \tilde{f}_n, \phi_k \rangle| + \sum_{l \ge 0} 2^{-l\left(\frac{d}{2}+1\right)} \sum_{k \in \mathbb{Z}^d} |\langle f - \tilde{f}_n, \psi_{lk} \rangle|, \tag{4.22}$$

where the constant depends only on the wavelet basis.

Remark 11. *Let $\hat{f}_n$ be some estimator of f, not necessarily with total mass 1. We obtain an estimator which integrates to 1 almost surely, which we call $\tilde{f}_n$, by renormalising the first wavelet layer of $\hat{f}_n$, that is, renormalising $\hat{f}_0 := K_0(\hat{f}_n)$. Then we set*

$$\tilde{f}_n = \frac{\hat{f}_0}{\int \hat{f}_0(x)\, \mathrm{d}x} + \sum_{l \ge 0} \sum_{k \in \mathbb{Z}^d} \langle \hat{f}_n, \psi_{lk} \rangle \psi_{lk}.$$

Note that while one can perform this procedure for any estimator $\hat{f}_n$, it is particularly simple for wavelet-based estimators. Assuming L_1-consistency of $\hat{f}_n$, $\hat{f}_0 \to K_0(f)$ and thus

$K_0(\tilde{f}_n) \to K_0(f)$ in L_1. Moreover, for the wavelet estimators we use below, this convergence occurs very fast, at the rate $n^{-\frac{S}{2S+d}}$, where S is the regularity of the wavelet basis. Thus it suffices to consider the un-normalised estimator $\hat{f}_n$ in the decomposition (4.22) whenever $s \leq S-1$, which we do in the sequel.

We first establish an upper bound for the estimation rate over the class $\mathcal{G}(s)$.

Theorem 20. *For any $s \geq 0$, there exists an estimator $\hat{f}_n$ such that for all sufficiently large n,*

$$\sup_{f \in \mathcal{G}(s)} E_f W_1(\hat{f}_n, f) \lesssim \begin{cases} (\log n)^{\frac{\gamma d}{2}+1} n^{-1/2}, & d = 2, \\ (\log n)^{\frac{\gamma d}{2}} n^{-\frac{s+1}{2s+d}}, & d \geq 3. \end{cases}$$

where γ is a constant depending on α and β only, and the constant depends on the parameters of the class $\mathcal{G}(s)$ and the wavelet basis. For $d = 1$, the empirical measure P_n satisifies

$$\sup_{f \in \mathcal{G}(s)} E_f W_1(P_n, P_f) \lesssim n^{-1/2}.$$

Remark 12. *These rates are sharp up to a logarithmic factor so long as L is sufficiently large: one uses a reduction to a multiple testing problem as in the proof of the lower bounds in Proposition 6, and then uses an analogous collection of well-separated densities defined on some common compact set. For large enough L, the compact support ensures that these densities have suitable exponential moments and so belong to $\mathcal{G}(s)$.*

Remark 13. *An inspection of the proof reveals that in fact it suffices to assume a suitable polynomial moment, depending on s; however, for convenience we assume an exponential moment which works for all $s \geq 0$.*

The proofs of Proposition 8 and Theorem 20 are given in Appendix 4.7. The estimator $\hat{f}_n$ is simply a wavelet projection estimator which is zero outside of a growing compact set; the risk outside of the compact is controlled using the moment assumption.

As in the case of $\mathbb{T}^d$, we require an adaptive estimator.

Theorem 21. *Let $d \geq 2$, and let $\gamma > 0$ be as in Theorem 20. Then there exists an estimator $\hat{f}_n$ of f such that for all $n \geq n_0(B)$ and all $s \geq 0$,*

$$\sup_{f \in \mathcal{G}(s)} E_f W_1(\hat{f}_n, f) \lesssim (\log n)^{\frac{\gamma d}{2}} \left(\frac{n}{\log n} \right)^{-\frac{s+1}{2s+d}},$$

where the constant depends on the parameters of the class $\mathcal{G}(s)$ and the wavelet basis.

The definition of $\hat{f}_n$ and proof of Theorem 21 are given in Appendix 4.7.

4.4.3 Construction of Confidence Sets

Let us now concretely state the two-class adaptation problem we wish to solve. Fix two smoothnesses $s > r \geq 0$ and consider the model $\mathcal{G}(r) = \mathcal{G}(r) \cup \mathcal{G}(s)$. Given $\alpha \in (0,1)$, we seek a confidence set C_n which has honest coverage at level $1-\alpha$, that is, for all n sufficiently large,

$$\inf_{f \in \mathcal{G}(r)} P_f(f \in C_n) \geq 1 - \alpha, \tag{4.23}$$

as well as the two diameter shrinkage conditions: for all $\alpha' > 0$ there exists a constant $K = K(\alpha') > 0$ such that

$$\sup_{f \in \mathcal{G}(r)} P_f(|C_n|_{W_1} > K R_n(r)) \leq \alpha', \tag{4.24}$$

$$\sup_{f \in \mathcal{G}(s)} P_f(|C_n|_{W_1} > K R_n(s)) \leq \alpha', \tag{4.25}$$

where $R_n(r)$ and $R_n(s)$ equal the convergence rates in Theorem 20 up to a poly-logarithmic factor.

As discussed previously, the $d = 1$ and $d = 2$ cases are straightforward given the existence of the estimator from Theorem 21, since here the convergence rates do not depend on the smoothness r. We thus restrict our attention to the case $d \geq 3$.

Let $X_1, \ldots, X_{2n}$ be an i.i.d. sample from the unknown $f \in \mathcal{G}(r)$. We split the sample as before into two equal halves, indexed by $\mathcal{S}^1 = 1, \ldots, n$ and $\mathcal{S}^2 = \{n+1, \ldots, 2n\}$, and denote by $P^{(i)}, E^{(i)}$ probabilities and expectations taken over $\mathcal{S}^i$. We wish to construct a confidence set via risk estimation, centred at the adaptive estimator $\hat{f}_n$ from Theorem 21, which we compute using $\mathcal{S}^1$. Proposition 8 provides a natural upper bound for $W_1(f, \hat{f}_n)^2$ which we then decompose into several terms. Define the thresholds $\kappa_{-1n} = \kappa_{0n} \simeq (\log n)^\gamma$, $\kappa_{ln} = 2^l \kappa_{0n}$ for γ chosen as in Theorem 20. Applying the Cauchy-Schwarz inequality several times, we obtain the bound

$$\begin{aligned} W_1(f, \hat{f}_n)^2 \leq 3 \Bigg(& (\log n)^{\gamma(d+2)} \left[\sum_{\|k\|_\infty \leq \kappa_{-1n}} \langle f - \hat{f}_n, \phi_k \rangle^2 + j \sum_{l<j} 2^{-2l} \sum_{\|k\|_\infty \leq \kappa_{ln}} \langle f - \hat{f}_n, \psi_{lk} \rangle^2 \right] \\ & + \left[\sum_{l \geq j} 2^{-l\left(\frac{d}{2}+1\right)} \sum_{\|k\|_\infty \leq \kappa_{ln}} |\langle f - \hat{f}_n, \psi_{lk} \rangle| \right]^2 \\ & + \left[\sum_{\|k\|_\infty > \kappa_{-1n}} \|k\| |\langle f, \phi_k \rangle| + \sum_{l \geq 0} 2^{-l\left(\frac{d}{2}+1\right)} \sum_{\|k\|_\infty > \kappa_{ln}} |\langle f, \psi_{lk} \rangle| \right]^2 \Bigg). \end{aligned} \tag{4.26}$$

The final term is controlled using the moment assumption on $f \in \mathcal{G}(r)$; indeed, from the proof of Theorem 20 we have that for all $f \in \mathcal{G}(r)$, this term is bounded above by

$$\Delta_n := C(d) L^2 (\log n)^{2\gamma} n^{-1}, \tag{4.27}$$

where $C(d)$ is a constant depending only on d and the wavelet basis.

We next consider the remaining terms in (4.26). We introduce pseudo-distances $\tilde{W}^{(n,j)}(f, g)$ defined as

$$\begin{aligned} \tilde{W}^{(n,j)}(f, g) = & \left[\sum_{\|k\|_\infty \leq \kappa_{-1n}} \langle f - g, \phi_k \rangle^2 + j \sum_{l<j} 2^{-2l} \sum_{\|k\|_\infty \leq \kappa_{ln}} \langle f - g, \psi_{lk} \rangle^2 \right]^{1/2} \\ & + \sum_{l \geq j} 2^{-l\left(\frac{d}{2}+1\right)} \sum_{\|k\|_\infty \leq \kappa_{ln}} |\langle f - g, \psi_{lk} \rangle|. \end{aligned} \tag{4.28}$$

Observe that for $f, g \in \mathcal{G}(r)$,

$$W_1(f, g) \leq \sqrt{3(\log n)^{\gamma(d+2)}} \cdot \tilde{W}^{(n,j)}(f, g) + \sqrt{3\Delta_n};$$

this is true uniformly over $r \geq 0$. Since $\sqrt{\Delta_n}$ converges (up to a logarithmic factor) at the parametric rate, this means that any diameter shrinkage condition with respect to $\tilde{W}^{(n,j)}$ provides an analogous shrinkage condition for W_1, with only a slightly worse rate. Moreover, the first part of $\tilde{W}^{(n,j)}(f,g)$ is well-suited to estimation using a U-statistic. To this end, define the U-statistic

$$V_{n,j} = V_{n,j}(\hat{f}_n) := \frac{2}{n(n-1)} \sum_{i<i',i,i' \in \mathcal{S}^2} \left[\sum_{\|k\|_\infty \leq \kappa_{-1n}} \left(\phi_k(X_i) - \langle \hat{f}_n, \phi_k \rangle\right)\left(\phi_k(X_{i'}) - \langle \hat{f}_n, \phi_k \rangle\right) + j \sum_{l<j} 2^{-2l} \sum_{\|k\|_\infty \leq \kappa_{ln}} \left(\psi_{lk}(X_i) - \langle \hat{f}_n, \psi_{lk} \rangle\right)\left(\psi_{lk}(X_{i'}) - \langle \hat{f}_n, \psi_{lk} \rangle\right) \right]. \tag{4.29}$$

Clearly we have that $E_f^{(2)} V_{n,j}$ is equal to the square of the first term in (4.28) with $f, \hat{f}_n$ in place of f, g. Analogously to Lemma 31, one shows that $V_{n,j}$ has small variance.

Lemma 33. *For $f \in L_\infty(\mathbb{R}^d)$, we have that, for some constant C_d depending only on d and the wavelet basis,*

$$\begin{aligned}
\mathrm{Var}_f^{(2)}(V_{n,j}) &\leq \frac{C_d}{2} \left(\frac{j^2 \|f\|_\infty^2 (\log n)^{\gamma d}}{n(n-1)} \sum_{l<j} 2^{l(d-4)} \right. \\
&\qquad \left. \ldots + \frac{\|f\|_\infty}{n} \left[\sum_{\|k\|_\infty \leq \kappa_{-1n}} \langle f - \hat{f}_n, \phi_k \rangle^2 + j^2 \sum_{l<j} 2^{-4l} \sum_{\|k\|_\infty \leq \kappa_{ln}} \langle f - \hat{f}_n, \psi_{lk} \rangle^2 \right] \right) \\
&\leq C_d \left(\frac{j^2 \|f\|_\infty^2 (\log n)^{\gamma d}}{n(n-1)} \sum_{l<j} 2^{l(d-4)} + \tilde{W}^{(n,j)}(f, \hat{f}_n)^2 \right) \\
&=: \lambda_{j,n}^2(f).
\end{aligned}$$

For the second part of $\tilde{W}^{(n,j)}(f, \hat{f}_n)$, we use the concentration arguments from the proof of Theorem 21 to show that this term is suitably small with high probability uniformly over $f \in \mathcal{G}(r)$.

Given a sequence (j_n), we write $\tilde{W}^{(n)}$ for $\tilde{W}^{(n,j_n)}$, and V_{j_n} for V_{n,j_n}.

Theorem 22. *Let $d \geq 3$. Fix $B \geq 1, M > 0, \alpha, \beta, L > 0, 1 \leq q \leq \infty$, and $s > r \geq 0$. Let $\gamma \geq 1$ be as in Theorem 20. If $d > 4$, assume additionally that $s \leq \frac{2d-4}{d-4} r + \frac{d}{d-4}$. Fix $\alpha \in (0,1)$. Consider the confidence set based on a sample of size $2n$ given by*

$$C_n = \left\{ g \in \mathcal{G}(r) : \tilde{W}^{(n)}(g, \hat{f}_n) \leq C(d) \sqrt{z_\alpha \lambda_{n,j_n}(g) + V_{j_n} + G_{j_n}} \right\} \tag{4.30}$$

where $\hat{f}_n$ is computed on $\mathcal{S}^1$, V_{j_n} is computed on $\mathcal{S}^2$, $C(d)$ is a constant depending on d and the wavelet basis, and:

- *$\lambda_{n,j_n}(g)$ is as in Lemma 33;*
- *j_n is such that $2^{j_n} \simeq \left(\frac{n}{\log n}\right)^{\frac{1}{2r+d/2}}$;*
- *$G_{j_n} = (\log n)^{\gamma d+1} 2^{-2j_n(r+1)}$;*

- $z_\alpha = (\alpha/2)^{-1/2}$.

Then for all $n \geq n_0(B)$, C_n satisfies (4.23), as well as (4.24) and (4.25) for a suitable constant $K > 0$ with the rates

$$R_n(r) = (\log n)^{\gamma(d+1)} \left(\frac{n}{\log n}\right)^{-\frac{r+1}{2r+d}}, \quad R_n(s) = (\log n)^{\gamma(d+1)} \left(\frac{n}{\log n}\right)^{-\frac{s+1}{2s+d}}.$$

In particular, C_n is a near-optimal adaptive W_1 confidence set over $\mathcal{F}(s) \cup \mathcal{F}(r)$.

The proof is almost identical to that of Theorem 18; a more detailed argument can be found in Appendix 4.7. In particular, this proves statements (i) and (ii) of Theorem 16.

4.4.4 Non-Existence of Confidence Sets

We now turn to the non-existence result (iii) in Theorem 16, a consequence of Lemma 32 (which holds in a general decision theoretic framework). We therefore require a lower bound on the minimax separation rate in the testing problem

$$H_0 : f \in \mathcal{G}(s) \quad \text{vs.} \quad H_1 : f \in \tilde{\mathcal{G}}(r, \rho), \tag{4.31}$$

where the separated alternative $\tilde{\mathcal{G}}(r, \rho)$ is defined analogously to before:

$$\tilde{\mathcal{G}}(r, \rho) := \{f \in \mathcal{G}(r) : W_1(f, \mathcal{G}(s)) \geq \rho\}.$$

Theorem 23. *Assume that $d > 4$ and $s > r \geq 0$. Let ρ_n^* be the minimax rate of testing for the problem (4.31). Then, for L sufficiently large in (4.21), there exist a constant $c > 0$ depending on the parameters of the class $\mathcal{G}(s)$ and the wavelet basis, and $n_0 = n_0(B, M)$ such that for all $n \geq n_0$,*

$$\rho_n^* \geq c n^{-\frac{r+1}{2r+d/2}}.$$

Also, (4.18) holds for any $\beta < 1$.

The proof is given in Appendix 4.7, and is similar to the proof of Theorem 19. As before, this implies statement (iii) of Theorem 16.

4.5 Wavelets and Besov Spaces

Here we introduce the wavelet bases we use, and define the various norms and spaces used in our analysis.

4.5.1 Wavelet Bases of $\mathbb{R}^d$ and $\mathbb{T}^d$

Let $S \in \mathbb{N}$. We begin with an S-regular wavelet basis of $L_2(\mathbb{R})$ generated by scaling function Φ and wavelet function Ψ,

$$\left\{\Phi_k = \Phi(\cdot - k), \Psi_{lk} = 2^{l/2}\Psi(2^l(\cdot) - k) : l \geq 0, k \in \mathbb{Z}\right\}.$$

Concretely, we take sufficiently regular Daubechies wavelets: see [45], [69, Chapter 4] , [119] for details. Such a wavelet basis has the following properties:

- Φ, Ψ are in $C^S(\mathbb{R})$, $\int_{\mathbb{R}} \Phi = 1$, and Ψ is orthogonal to polynomials of degree $< S$.
- $\|\sum_k |\Phi_k|\|_\infty \lesssim 1$, and$\|\sum_k |\Psi_{lk}|\|_\infty \lesssim 2^{l/2}$ for a constant depending only on Ψ.
- Letting $V_j = \operatorname{span}(\Phi_k, \Psi_{lk} : l < j)$, for any $f \in V_j$ the following Bernstein estimate holds:
$$\|\nabla f\|_p \lesssim 2^j \|f\|_p ,$$
for a constant depending only on the wavelet basis.
- Φ, Ψ are compactly supported.

We then form a tensor product basis of $L_2(\mathbb{R}^d)$ as follows. Let $\mathcal{I} = \{0,1\}^d \setminus \{0\}$. Define

$$\phi(x) = \Phi(x_1) \cdots \Phi(x_d), \quad x \in \mathbb{R}^d$$

and, writing $\Psi^0 = \Phi, \Psi^1 = \Psi$,

$$\psi^\iota = \Psi^{\iota_1}(x_1) \cdots \Psi^{\iota_d}(x_d), \quad \iota \in \mathcal{I}.$$

Then ([69, Section 4.3.6])

$$\left\{\phi_k = \phi(\cdot - k), \psi^\iota_{lk} = 2^{ld/2}\psi^\iota(2^l x - k) : l \geq 0, k \in \mathbb{Z}^d, \iota \in \mathcal{I}\right\}$$

defines a wavelet basis of $L_2(\mathbb{R}^d)$. We omit ι from our notation and simply write ψ_{lk} with k now implicitly taking values in $\mathbb{Z}^d \times \mathcal{I}$; any sum over k is to be understood as over all $\iota \in \mathcal{I}$ as well.

i) ϕ, ψ are in $C^S(\mathbb{T}^d)$, $\int_{\mathbb{R}^d} \phi = 1$, and ψ is orthogonal to polynomials of degree $< S$.

ii) $\|\sum_k |\phi_k|\|_\infty \lesssim 1$, and$\|\sum_k |\psi_{lk}|\|_\infty \lesssim 2^{ld/2}$ for a constant depending only on ψ.

iii) ϕ, ψ are compactly supported.

These properties follow elementarily from the previously stated properties of Φ and Ψ. Property 3) is used crucially in our analysis on $\mathbb{R}^d$. Notably, this precludes certain common choices of wavelet basis, such as the Meyer basis.

These properties imply the following relationship between L_p-norms of functions and the ℓ_p-norms of their wavelet coefficients (by an abuse of notation we denote both of these norms by $\|\cdot\|_p$).

Lemma 34. *For any $l \geq 0$, any $p \in [1, \infty]$ and any $c \in \mathbb{R}^{\mathbb{Z}^d}$, we have that*

$$\left\| \sum_{k \in \mathbb{Z}^d} c_k \psi_{lk} \right\|_p \simeq 2^{ld(1/2 - 1/p)} \|c\|_p,$$

where the constants depend on ψ and p only.

When working on $\mathbb{T}^d$, we use the tensor product wavelet basis induced by the periodisations of Φ, Ψ; see [69, Section 4.3.4] for details. This produces a basis of $L_2(\mathbb{T}^d)$ with the following properties:

i) $\psi(x) = \prod_{i=1}^d \psi^{(i)}(x_i)$ for some univariate functions $\psi^{(i)}$.

ii) Setting $\psi_{lk}(\cdot) = 2^{ld/2}\psi(\cdot - 2^{-l}k)$ for $l \geq 0, k \in \mathbb{Z}^d \cap [0, 2^l)^d$, the set

$$\left\{\phi, \psi_{lk} : l \geq 0, k \in \mathbb{Z}^d \cap [0, 2^l)^d\right\}$$

forms an orthonormal basis of $L_2(\mathbb{T}^d)$. By an abuse of notation, we re-index in k such that $k \in \mathbb{Z}$ varies between $0 \leq k < 2^{ld}$.

iii) ψ is in $C^S(\mathbb{T}^d)$, and is orthogonal to polynomials of degree $< S$.

iv) $\|\sum_k |\psi_{lk}|\|_\infty \lesssim 2^{ld/2}$, for a constant depending only on ψ.

v) Letting $V_j = \operatorname{span}(\phi, \psi_{lk} : l < j)$, for any $f \in V_j$ the following Bernstein estimate holds:

$$\|\nabla f\|_p \lesssim 2^j \|f\|_p ,$$

for a constant depending only on the wavelet basis.

Again, these are basic consequences of properties of Φ, Ψ, and enable the proof of Proposition 5; compare to Appendix C of [171].

4.5.2 Besov Spaces

In this section, we let (ϕ_k, ψ_{lk}) denote either the S-regular tensor product Daubechies wavelet basis of $L_2(\mathbb{R}^d)$, or the S-regular tensor product periodised Daubechies wavelet basis of $L_2(\mathbb{T}^d)$. It should be understood that any summation is over the full range of indices, for example $\sum_k \psi_{lk}$ denotes $\sum_{k\in\mathbb{Z}^d} \psi_{lk}$ in the $\mathbb{R}^d$ case and $\sum_{k=0}^{2^{ld}-1} \psi_{lk}$ in the $\mathbb{T}^d$ case. We further let $\mathcal{D}$ be either the class of tempered distributions on $\mathbb{R}^d$, or the class of periodic tempered distributions on $\mathbb{T}^d$.

Let $1 \leq p \leq \infty$, $1 \leq q \leq \infty$, $s \in \mathbb{R}, s < S$. For $f \in \mathcal{D}$, we define the Besov norm

$$\|f\|_{B^s_{pq}} = \|\langle f, \phi_\cdot\rangle\|_p + \left(\sum_{l\geq 0}\left[2^{ls}2^{ld\left(\frac{1}{2}-\frac{1}{p}\right)} \|\langle f, \psi_{l\cdot}\rangle\|_p\right]^q\right)^{1/q}, \tag{4.32}$$

where $\|\cdot\|_p$ is the ℓ_p-norm. When $q = \infty$, the norm is defined as

$$\|f\|_{B^s_{p\infty}} = \|\langle f, \phi_\cdot\rangle\|_p + \sup_{l\geq 0} 2^{ls}2^{ld\left(\frac{1}{2}-\frac{1}{p}\right)} \|\langle f, \psi_{l\cdot}\rangle\|_p . \tag{4.33}$$

We then define the corresponding Besov space B^s_{pq} as

$$B^s_{pq} = \left\{f \in \mathcal{D} : \|f\|_{B^s_{pq}} < \infty\right\}. \tag{4.34}$$

We will write $B^s_{pq}(\mathbb{R}^d)$ or $B^s_{pq}(\mathbb{T}^d)$ to remove any ambiguity over the choice of domain, whenever it arises.

The definition of B^s_{pq} is independent of the wavelet basis used, that is, using a different (sufficiently regular) basis in the definition (4.32) produces an equivalent norm. Moreover, using a C^∞ basis such as the Meyer basis enables us to define B^s_{pq} concurrently for all $s \in \mathbb{R}$.

4.5.3 The Case of the Unit Cube

We can also define a 'boundary-corrected' wavelet basis of $L_2([0,1]^d)$ based on Φ, Ψ as in [43]; see also [69, Section 4.3.5]. Such a basis possesses completely analogous properties to properties 1)-5) of the periodised basis of $L_2(\mathbb{T}^d)$; moreover, all Besov spaces defined on $\mathbb{T}^d$ are defined on the unit cube $[0,1]^d$ by replacing the periodised wavelet basis with the boundary-corrected wavelet basis (as used in [170]). Thus all of our results for $\mathbb{T}^d$ hold also for the case of $[0,1]^d$.

4.6 Proofs for Section 4.3

We first give the definition of our adaptive estimator. The estimator is based on the empirical wavelet coefficients, defined as

$$\hat{f}_{lk} := \frac{1}{n}\sum_{i=1}^{n} \psi_{lk}(X_i).$$

We also write $f_{l\cdot}$ and $\hat{f}_{l\cdot}$ for the vectors of coefficients $(f_{lk} : 0 \le k < 2^{ld})$ and $(\hat{f}_{lk} : 0 \le k < 2^{ld})$ respectively.

Next, define the truncation point $l_{\max}$ such that

$$2^{l_{\max}} \simeq \left(\frac{n}{\log n}\right)^{1/d},$$

and for $0 \le l \le l_{\max}$, define the thresholds

$$\tau_l := \tau 2^{\frac{ld}{2}} \left(\frac{\log n}{n}\right)^{1/2},$$

for some $\tau > 0$ to be chosen below, depending only on B, d, M and the wavelet basis. We then define

$$\hat{f}_n := 1 + \sum_{l=0}^{l_{\max}} \mathbb{1}_{\{\|\hat{f}_{l\cdot}\|_2^2 > \tau_l^2\}} \sum_{k=0}^{2^{ld}-1} \hat{f}_{lk}\psi_{lk}. \tag{4.35}$$

To prove Theorem 17, we must first collect some results on the expectation and concentration of the empirical wavelet coefficients $\hat{f}_{lk}$.

Lemma 35. *Let $f \in \mathcal{F}(s)$ and let $\hat{f}_{lk}$ be the empirical wavelet coefficients of f based on a sample of n observations. Then for every $t \ge 2$ there exists a constant C_t depending only on t such that for all $l \ge 0$ satisfying $2^{ld} \le n$,*

$$E\left|\hat{f}_{lk} - f_{lk}\right|^t \le C_t M \|\psi\|_\infty^{t-2} n^{-t/2}.$$

For $t = 2$, the proof is immediate from the i.i.d. assumption on the data, the orthonormality of the wavelets and the bound $\|f\|_\infty \le M$. For $t > 2$, the result follows from the $t = 2$ case and Hoffmann-Jørgensen's inequality ([69, Theorem 3.1.22], [79]). We also require a concentration result for the $\hat{f}_{lk}$; for this we use Bernstein's inequality ([69, Theorem 3.1.7]).

Proposition 9. *[Bernstein's Inequality] Let $Y_1, \dots, Y_n$ be independent centred random variables which are almost surely bounded by $c > 0$ in absolute value. Let $\sigma^2 = n^{-1}\sum_{i=1}^n EY_i^2$ and $S_n = \sum_{i=1}^n Y_i$. Then for all $u \ge 0$,*

$$P(|S_n| > u) \le 2\exp\left(-\frac{u^2}{2n\sigma^2 + \frac{2cu}{3}}\right).$$

For fixed l, k and $f \in \mathcal{F}(s)$, the random variables $(\psi_{lk}(X_i) - f_{lk})$ are i.i.d., centred, bounded by $2^{ld/2}\|\psi\|_\infty =: c_l$, and have variance bounded by M. Thus from Bernstein's inequality, we deduce that

$$P_f\left(|\hat{f}_{lk} - f_{lk}| > u\right) \leq 2\exp\left(-\frac{nu^2}{2M + \frac{2c_l u}{3}}\right). \tag{4.36}$$

We also need a result on wavelet approximations in the $H^{-1,\delta}$ norm to control bias terms. The following lemma about the error of j-level approximations to Besov functions is standard; see Propositions 4.3.8 and 4.3.14 in [69], for instance.

Lemma 36. *Let $0 \leq s < S$ and $1 \leq q \leq \infty$, $\delta \in \mathbb{R}$. Then for $f \in B^s_{2q}$, we have that*

$$\|K_j(f) - f\|_{H^{-1,\delta}} \leq C \sup_{l \geq j}\left(2^{-l(s+1)} l^\delta\right) \|f\|_{B^s_{2q}}, \tag{4.37}$$

where the constant C depends only on the wavelet basis. In particular, for $j \geq 1 \vee \frac{\delta}{s+1}$, we have that

$$\|K_j(f) - f\|_{H^{-1,\delta}} \leq C 2^{-j(s+1)} j^\delta \|f\|_{B^s_{2q}}$$

Proof of Theorem 17. Fix $f \in \mathcal{F}(s)$. Define $l_n(s)$ such that

$$2^{l_n(s)} \simeq B^{\frac{1}{s}}\left(\frac{n}{\log n}\right)^{\frac{1}{2s+d}};$$

for all sufficiently large n depending on B, we have that $l_n(s) < l_{\max}$. We then decompose the risk as follows:

$$\begin{aligned}
\|f - \hat{f}_n\|^2_{H^{-1,\delta}} &= \sum_{l=0}^{l_n(s)} 2^{-2l}(l \vee 1)^{2\delta} \|\langle f - \hat{f}_n, \psi_{l\cdot}\rangle\|_2^2 + \sum_{l=l_n(s)+1}^{l_{\max}} 2^{-2l} l^{2\delta} \|\langle f - \hat{f}_n, \psi_{l\cdot}\rangle\|_2^2 \\
&\quad + \sum_{l > l_{\max}} 2^{-2l} l^{2\delta} \|\langle f, \psi_{l\cdot}\rangle\|_2^2 \\
&=: I + II + III.
\end{aligned} \tag{4.38}$$

This is a bias-stochastic decomposition, where we have further divided the stochastic term into terms I and II.

We first deal with the bias term III: a direct application of Lemma 36 gives

$$\begin{aligned}
III &= \|K_{l_{\max}}(f) - f\|^2_{H^{-1,\delta}} \\
&\lesssim l_{\max}^{2\delta} 2^{-2l_{\max}(s+1)} \\
&= o\left((\log n)^{2\delta}\left(\frac{n}{\log n}\right)^{-\frac{2(s+1)}{2s+d}}\right)
\end{aligned}$$

for a constant depending on B and the wavelet basis.

Next, we deal with term I. For any $l \geq 0$, by the triangle inequality we have that

$$\|\langle f - \hat{f}_n, \psi_{l\cdot}\rangle\|_2 \leq \|f_{l\cdot} - \hat{f}_{l\cdot}\|_2 + \|\hat{f}_{l\cdot}\|_2 \mathbb{1}_{\{\|\hat{f}_{l\cdot}\|_2 \leq \tau_l\}} \leq \|f_{l\cdot} - \hat{f}_{l\cdot}\|_2 + \tau 2^{ld/2}\sqrt{\frac{\log n}{n}}.$$

Using Lemma 35 to control the expectation of the square of the first term, we see that

$$E_f(I) \lesssim \sum_{l=0}^{l_n(s)} 2^{-2l}(l \vee 1)^{2\delta} \left[2^{ld} n^{-1} + \tau^2 2^{ld} \frac{\log n}{n}\right]$$
$$\lesssim \tau^2 \frac{\log n}{n} (l_n(s))^{2\delta} \sum_{l=0}^{l_n(s)} 2^{l(d-2)},$$

for n large enough. Note that $l_n(s) \lesssim \log n$. Thus when $d = 2$, the sum contributes at most some power of $\log n$, and so $E_f(I)$ is clearly sufficiently small. For $d > 2$, the final term dominates the sum and so using the definition of $l_n(s)$,

$$E_f(I) \lesssim \tau^2 (\log n)^{2\delta} \left(\frac{n}{\log n}\right)^{-\frac{2(s+1)}{2s+d}}$$

as required.

Lastly, we must analyse term II. Since we consider resolution levels $l > l_n(s)$, we have that

$$\|f_{l\cdot}\|_2 \leq B 2^{-ls} < B 2^{-l_n(s)s} \simeq \left(\frac{n}{\log n}\right)^{-\frac{s}{2s+d}},$$

for a constant depending only on B. Moreover,

$$\tau_l = \tau 2^{ld/2} \left(\frac{n}{\log n}\right)^{-1/2} > \tau 2^{l_n(s)d/2} \left(\frac{n}{\log n}\right)^{-1/2} \geq \tau \left(\frac{n}{\log n}\right)^{-\frac{s}{2s+d}},$$

and so for τ chosen sufficiently large depending only on B, we have that $\|f_{l\cdot}\|_2 \leq \tau_l/2$. Define events

$$A_{l,n} := \left\{\|\hat{f}_{l\cdot}\|_2 \leq \tau_l\right\}, \quad l_n(s) < l \leq l_{\max}.$$

Then by the above observations, the triangle inequality, a union bound and the bound (4.36), we have that

$$\begin{aligned} P_f(A_{l,n}^c) &\leq P_f(\|\hat{f}_{l\cdot} - f_{l\cdot}\|_2 > \tau_l/2) \\ &\leq \sum_{k=0}^{2^{ld}-1} P_f\left(|\hat{f}_{lk} - f_{lk}| > \frac{\tau}{2}\sqrt{\frac{\log n}{n}}\right) \\ &\leq 2^{ld} \cdot 2\exp\left(-\frac{\tau^2 n \log n/4}{2Mn + \tau c_l \sqrt{n \log n}/3}\right) \\ &\lesssim \frac{n}{\log n} \exp\left(-C\tau \log n\right), \end{aligned} \tag{4.39}$$

for τ large enough depending on M and the wavelet basis, as $l \leq l_{\max}$ and so $2^l \leq (n/\log n)^{1/d}$. Here, C is an absolute constant. Note that on the event $A_{l,n}^c$, $\langle \hat{f}_n, \psi_{lk}\rangle = \hat{f}_{lk}$, whereas on $A_{l,n}$, $\langle \hat{f}_n, \psi_{lk}\rangle = 0$. Thus for $l_n(s) < l \leq l_{\max}$,

$$E_f \|\langle \hat{f}_n - f, \psi_{l\cdot}\rangle\|_2^2 \mathbb{1}_{A_{l,n}} \leq \|\langle f, \psi_{l,\cdot}\rangle\|_2^2 \lesssim 2^{-2ls} \tag{4.40}$$

for some constant depending on B, using (4.3). Next, using Cauchy-Schwarz in conjunction with (4.39) and Lemma 35,

$$E_f \|\langle \hat{f}_n^T - f, \psi_{l\cdot}\rangle\|_2^2 \mathbb{1}_{A_{l,n}^c} = \sum_{k=0}^{2^{ld}-1} E_f |\hat{f}_{lk} - f_{lk}|^2 \mathbb{1}_{A_{l,n}^c}$$

$$\leq \sum_{k=0}^{2^{ld}-1} \left(E_f|\hat{f}_{lk} - f_{lk}|^4\right)^{1/2} \left(P_f(A_{l,n}^c)\right)^{1/2}$$
$$\lesssim 2^{ld}(n\log n)^{-1/2}n^{-C\tau/2}. \tag{4.41}$$

Combining the estimates (4.40) and (4.41), we may bound II as follows:

$$E_f(II) \lesssim \sum_{l=l_n(s)+1}^{l_{\max}} 2^{-2l}l^{2\delta}\left[2^{-2ls} + 2^{ld}(n\log n)^{-1/2}n^{-C\tau/2}\right]$$
$$\lesssim (\log n)^{2\delta}\left[2^{-2(s+1)l_n(s)} + (n\log n)^{-1/2}n^{-C\tau/2}\sum_{l\leq l_{\max}} 2^{l(d-2)}\right].$$

By the definition of $l_n(s)$, the first term is of the correct order. It remains to consider the second term. When $d = 2$ the sum contributes a logarithmic factor and so the second term is clearly sufficiently small. When $d > 2$, the sum is dominated by its final term and so the second term inside the brackets is of order

$$(n\log n)^{-1/2}n^{-C\tau/2}2^{l_{\max}(d-2)} \simeq \log n^{-1/2}n^{\frac{1}{2}-\frac{2}{d}-\frac{C\tau}{2}};$$

by choosing τ sufficiently large, we can make this term sufficiently small for all $s \geq 0$. This concludes the proof. □

We will also later require the following lemma, which gives control of the B_{2q}^s norm of the estimator $\hat{f}_n$.

Lemma 37. *Under the hypotheses of Theorem 17, given $\alpha \in (0,1)$ there exists $n_0 = n_0(\alpha)$ such that for all $n \geq n_0$ and any $f \in \mathcal{F}(s)$, with P_f-probability at least $1-\alpha$,*

$$\|\hat{f}_n\|_{B_{2q}^s} \lesssim B + \tau B^{d/2s},$$

where the constant depends on d, q only.

Proof. Let $l_n(s), A_{l,n}$ be as in the previous proof. Further define events $B_{l,n} = \{\|\hat{f}_{l\cdot} - f_{l\cdot}\|_2 \leq \tau_l\}$, and

$$A_n = \left(\bigcap_{0\leq l\leq l_n(s)} B_{l,n}\right) \bigcap \left(\bigcap_{l_n(s)<l\leq l_{\max}} A_{l,n}\right).$$

We have from (4.39), which holds with $B_{l,n}$ in place of $A_{l,n}$ when $l \leq l_n(s)$, and a union bound that

$$P_f(A_n^c) \lesssim l_{\max}\frac{n}{\log n}\exp\left(-C\tau\log n\right) \lesssim n\exp\left(-C\tau\log n\right)$$

and so by choosing $\tau > 0$ sufficiently large (independently of α), we can make this smaller than α for all sufficiently large n. Then on the event A_n, using $(a+b)^q \leq 2^{q-1}(a^q+b^q)$,

$$\|\hat{f}_n\|_{B_{2q}^s}^q = 1 + \sum_{l=0}^{l_{\max}} 2^{lqs}\mathbb{1}_{\{\|\hat{f}_{l\cdot}\|_2 > \tau_l\}}\|\hat{f}_{l\cdot}\|_2^q$$
$$\lesssim 1 + \sum_{l=0}^{l_n(s)} 2^{lqs}\|f_{l\cdot}\|_2^q + \sum_{l=0}^{l_n(s)} 2^{lqs}\|\hat{f}_{l\cdot} - f_{l\cdot}\|_2^q$$

$$
\begin{aligned}
&\leq \|f\|_{B^s_{2q}}^q + \sum_{l=0}^{l_n(s)} 2^{lqs}\tau_l^q \\
&= B^q + \tau^q \left(\frac{\log n}{n}\right)^{q/2} \sum_{l=0}^{l_n(s)} 2^{lq\left(\frac{d}{2}+s\right)} \\
&\lesssim B^q + \tau^q B^{dq/2s},
\end{aligned}
$$

by choice of $l_n(s)$, since the sum is dominated by its largest term. □

Proof of Lemma 31. The kernel of the U-statistic is

$$
R(x,y) = \sum_{l\leq j-1} 2^{-2l}(l\vee 1)^{2\delta} \sum_{k=0}^{2^{ld}-1} \left[(\psi_{lk}(x) - \langle \psi_{lk}, \tilde{f}_n\rangle)(\psi_{lk}(y) - \langle \psi_{lk}, \tilde{f}_n\rangle)\right]
$$

which is symmetric, and so has Hoeffding decomposition (see Section 11.4 of [162])

$$
\begin{aligned}
U_n(\tilde{f}_n) - E_f^{(2)} U_n(\tilde{f}_n) &= \frac{2}{n}\sum_{i\in\mathcal{S}^2} (\pi_1 R)(X_i) + \frac{2}{n(n-1)} \sum_{i<i',i,i'\in\mathcal{S}^2} (\pi_2 R)(X_i, X_{i'}) \\
&=: L_n + D_n,
\end{aligned} \tag{4.42}
$$

with linear kernel

$$
(\pi_1 R)(x) = \sum_{l\leq j-1} 2^{-2l}(l\vee 1)^{2\delta} \sum_{k=0}^{2^{ld}-1} \left[(\psi_{lk}(x) - \langle \psi_{lk}, f\rangle)\langle \psi_{lk}, f - \tilde{f}_n\rangle\right]
$$

and degenerate kernel

$$
(\pi_2 R)(x,y) = \sum_{l\leq j-1} 2^{-2l}(l\vee 1)^{2\delta} \sum_{k=0}^{2^{ld}-1} \left[(\psi_{lk}(x) - \langle \psi_{lk}, f\rangle)(\psi_{lk}(y) - \langle \psi_{lk}, f\rangle)\right].
$$

One checks that L_n and D_n are uncorrelated. It thus remains to bound their variances separately. For $\mathrm{Var}^{(2)}(L_n)$, we use the uncentred version of the kernel $\pi_1 R$ and orthonormality of the wavelet basis

$$
\begin{aligned}
\mathrm{Var}^{(2)}(L_n) &\leq \frac{4}{n}\int \left(\sum_{l\leq j-1} 2^{-2l}(l\vee 1)^{2\delta} \sum_{k=0}^{2^{ld}-1} \psi_{lk}(x)\langle \psi_{lk}, f - \tilde{f}_n\rangle\right)^2 f(x)\,\mathrm{d}x \\
&\leq \frac{4\|f\|_\infty}{n}\left(\max_{l\geq -1} 4^{-l}(1\vee l)^{2\delta}\right) \sum_{l\leq j-1} 2^{-2l}(l\vee 1)^{2\delta} \sum_{k=0}^{2^{ld}-1} \langle \psi_{lk}, f - \tilde{f}_n\rangle^2 \\
&= \frac{4\|f\|_\infty}{n}\left(\max_{l\geq -1} 4^{-l}(1\vee l)^{2\delta}\right) \|K_j(f - \tilde{f}_n)\|_{H^{-1,\delta}}^2.
\end{aligned}
$$

We next bound $\mathrm{Var}^{(2)}(D_n)$. By the degeneracy of the kernel, the summands are uncorrelated. So

$$
\begin{aligned}
\mathrm{Var}^{(2)}(D_n) &\leq E^{(2)}\left(\frac{2}{n(n-1)} \sum_{i<i',i,i'\in\mathcal{S}^2} (\pi_2 R)(X_i, X_{i'})\right)^2 \\
&\leq \frac{2}{n(n-1)} E_f^{(2)}\left(\sum_{l\leq j-1} 2^{-2l}(l\vee 1)^{2\delta} \sum_{k=0}^{2^{ld}-1} [\psi_{lk}(X_i)\psi_{lk}(X_{i'})]\right)^2
\end{aligned}
$$

$$\leq \frac{2\|f\|_\infty^2}{n(n-1)} \sum_{l \leq j-1} 2^{-4l}(l \vee 1)^{4\delta} \sum_{k=0}^{2^{ld}-1} \left(\int \psi_{lk}(x)^2 \, \mathrm{d}x \right)^2$$
$$= \frac{2\|f\|_\infty^2}{n(n-1)} \sum_{l \leq j-1} 2^{l(d-4)}(l \vee 1)^{4\delta},$$

using the orthonormality of the wavelet basis. Combining these two estimates concludes the proof. □

Proof of Theorem 18. We first establish the coverage condition (4.7). By Lemma 37, for all n sufficiently large we have with P_f-probability at least $1 - \alpha/2$ that $\hat{f}_n$ is in a B^s_{2q}-norm ball of constant radius. Thus for any $f \in \mathcal{F}(r)$, with P_f-probability at least $1 - \alpha/2$, for $n \geq n_0(B, \alpha)$ we have from (4.37) that

$$\|K_{j_n}(f - \hat{f}_n) - (f - \hat{f}_n)\|^2_{H^{-1,\delta}} \leq G(j_n).$$

By conditioning on this event, we have that

$$\begin{aligned}
P_f(f \in C_n) &= P_f\left(U_{n,j}(\hat{f}_n) - \|f - \hat{f}_n\|^2_{H^{-1,\delta}} \geq -G(j) - z_\alpha \kappa_{n,j,\delta}(f)\right) \\
&\geq \left(1 - \frac{\alpha}{2}\right) P_f^{(2)}\left(U_{n,j}(\hat{f}_n) - \|K_j(f - \hat{f}_n)\|^2_{H^{-1,\delta}} \geq -z_\alpha \kappa_{n,j,\delta}(f)\right) \\
&\geq \left(1 - \frac{\alpha}{2}\right)\left(1 - \frac{\mathrm{Var}_f^{(2)}(U_{n,j}(\hat{f}_n))}{(z_\alpha \kappa_{n,j,\delta}(f))^2}\right) \\
&\geq \left(1 - \frac{\alpha}{2}\right)^2 \\
&\geq 1 - \alpha
\end{aligned}$$

by Chebyshev's inequality and Lemma 31.

We now move on to checking the diameter shrinkage conditions (4.8) and (4.9). Writing $S_j := \sum_{l<j} 2^{l(d-4)}(l \vee 1)^{4\delta}$ and using the fact that for positive numbers a, b, $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, for $g \in \mathcal{F}(r)$ we have that $\kappa_{n,j_n,\delta}(g) \leq 2\sqrt{M} n^{-1/2} \|g - \hat{f}_n\|_{H^{-1,\delta}} + 2M\sqrt{S_{j_n}} n^{-1}$ and so $g \in C_n$ if and only if

$$\|g - \hat{f}_n\|_{H^{-1,\delta}} \leq \sqrt{z_\alpha \frac{2M}{n}\sqrt{S_{j_n}} + U_{j_n} + G(j_n)} + n^{-1/4}\sqrt{2z_\alpha\sqrt{M}}\sqrt{\|g - \hat{f}_n\|_{H^{-1,\delta}}}.$$

For positive numbers x, a, b, the inequality $x \leq b + a\sqrt{x}$ implies that $x \leq 2b + 2a^2$. Applying this inequality with the values $x = \|g - \hat{f}_n\|_{H^{-1,\delta}}, a = n^{-1/4}\sqrt{2z_\alpha\sqrt{M}}, b = \sqrt{z_\alpha \frac{2M}{n}\sqrt{S_{j_n}} + U_{j_n} + G(j_n)}$, and further using that for any positive numbers x, y we have that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$, one sees that the diameter of C_n is bounded by a multiple of

$$n^{-1/2} S_{j_n}^{1/4} + \sqrt{U_{j_n}} + \sqrt{G_{j_n}} + n^{-1/2}.$$

We consider each of these terms separately; note that the final term is always sufficiently small.

First, consider $G(j_n)$: this is deterministic, of order

$$G(j_n) \lesssim (\log n)^{1+2\delta}\left(\frac{n}{\log n}\right)^{-\frac{2(r+1)}{2r+d/2}} = o(R_n(s)^2) = o(R_n(r)^2).$$

(When $d \leq 4$ this is trivial; for $d > 4$, it necessitates the assumption on s.)

Next, $n^{-2}S_{j_n}$ is of order

$$n^{-2} \sum_{l \leq j_n - 1} 2^{l(d-4)} (l \vee 1)^{4\delta}.$$

When $d \leq 4$, this contributes at most a logarithmic factor in n times n^{-2}, so this is clearly $o(R_n(s)^4)$ and $o(R_n(r)^4)$. When $d > 4$, the final term dominates the sum and so the contribution is of order

$$(\log n)^{4\delta - \frac{d-4}{2r+d/2}} n^{-\frac{4(r+1)}{2r+d/2}} = O(R_n(s)^4) = o(R_n(r)^4),$$

again by the assumption on s.

Finally, since $\mathrm{Var}(U_{j_n}) \to 0$ as $n \to \infty$, we know that

$$U_{j_n} = O_P\left(E_f U_{j_n}\right) = O_P\left(E_f \|K_j(f - \hat{f}_n)\|^2_{H^{-1,\delta}}\right) = O_P\left(E_f \|f - \hat{f}_n\|^2_{H^{-1,\delta}}\right).$$

As $\hat{f}_n$ converges at the rates $R_n(s)$ and $R_n(r)$ uniformly over $\mathcal{F}(s)$ and $\mathcal{F}(r)$ respectively, U_{j_n} is of the correct order in probability in both cases. This concludes the proof. □

Proof of Theorem 19. For some sequence $L_n \to \infty$, to be defined below, and any $\omega \in \{-1; 1\}^{\mathbb{Z}^d \cap [0, 2^{L_n})^d}$, we define for some $\epsilon > 0$,

$$f_{n,\omega} := 1 + \epsilon 2^{-L_n(r+d/2)} \sum_{k \in \mathbb{Z} \cap [0, 2^{L_n})^d} \omega_k \psi_{L_n,k}.$$

Provided that $B > 1$,

$$\begin{aligned} \|f_{n,\omega}\|_{B^r_{2q}} &= 1 + 2^{L_n r} \left(\sum_{k \in \mathbb{Z} \cap [0, 2^{L_n})^d} |\langle f_{n,\omega}, \psi_{L_n,k} \rangle|^2 \right)^{1/2} \\ &= 1 + \epsilon 2^{L_n r} 2^{-L_n(r+d/2)} 2^{dL_n/2} \\ &= 1 + \epsilon, \end{aligned}$$

ensuring that $f_{n,\omega}$ is in the $\|\cdot\|_{B^2_{2q}}$-Besov ball of radius B for ϵ small enough. Also, $\int_{\mathbb{T}^d} f_{n,\omega}(t)dt = 1$ and, as the tensor product wavelet basis is assumed to be $S-$regular (cf. Appendix 4.5),

$$\left\| \sum_k |\psi_{L_n,k}| \right\|_\infty \lesssim 2^{dL_n/2},$$

for some constant depending on the basis only. Therefore,

$$\|f_{n,\omega} - 1\|_\infty \leq \epsilon c 2^{-rL_n},$$

so that, for any $M > 1 \geq m > 0$, $f_{n,\omega} \in \mathcal{F}(r)$ for n large enough (or ϵ small enough if $r = 0$). Finally, for any $\rho_n = o\left(n^{-\frac{1+r}{2r+d/2}}\right)$, $f_{n,\omega} \in \tilde{\mathcal{F}}(r, \rho_n)$ if, for any $g \in \mathcal{F}(s)$, $W_2\left(f_{n,\omega}, g\right) \geq \rho_n$. By definition of $\mathcal{F}(r)$, $\mathcal{F}(s)$ and Proposition 5, we have, for n large enough

$$\begin{aligned} W_2\left(f_{n,\omega}, g\right)^2 &\gtrsim \|f_{n,\omega} - g\|^2_{B^{-1}_{2\infty}} \\ &\geq 2^{-2L_n} \left\|\langle f_{n,\omega} - g, \psi_{L_n,\cdot} \rangle\right\|_2^2 \end{aligned}$$

$$\geq 2^{-2L_n}\left[\left(\sum_{k=0}^{2^{L_n d}-1}|\langle f_{n,\omega},\psi_{L_n,k}\rangle|^2\right)^{1/2}-\left(\sum_{k=0}^{2^{L_n d}-1}|\langle g,\psi_{L_n,k}\rangle|^2\right)^{1/2}\right]^2$$
$$\geq 2^{-2L_n}\left[\epsilon 2^{-L_n r}-B2^{-L_n s}\right]^2$$
$$\geq \frac{\epsilon^2}{2}2^{-2L_n(1+r)}.$$

Therefore, if L_n^* is such that $2^{-2L_n^*(1+r)} \asymp n^{-2\frac{1+r}{2r+d/2}}$, it is possible to find $L_n > L_n^*$ such that $\rho_n^2 \leq \frac{\epsilon^2}{2}2^{-2L_n(1+r)} = o\left(n^{-2\frac{1+r}{2r+d/2}}\right)$. This choice ensures that, for any ω, $f_{n,\omega} \in \tilde{\mathcal{F}}(r,\rho_n)$. Note also that the density $f_0 := 1$ naturally belongs to $\mathcal{F}(s)$.

Re-index $\{-1;1\}^{\mathbb{Z}^d\cap\left[0,2^{L_n}\right)^d}$ as $\left\{\omega^{(i)}:\ i=1,\dots,2^{2^{dL_n}}\right\}$ and denote by P_i the distribution with Lebesgue density $f_i := f_{n,\omega^{(i)}}$, $Q := 2^{-2^{dL_n}}\sum_{i=1}^{2^{2^{L_n d}}} P_i$ and P_0 the distribution with density f_0. Then, with μ the Lebesgue measure and for any test Ψ_n,

$$\sup_{f\in\Sigma_0}\mathbb{E}_f\left[\Psi_n\right]+\sup_{f\in\Sigma(\rho_n)}\mathbb{E}_f\left[1-\Psi_n\right] \geq \mathbb{E}_{f_0}\left[\Psi_n\right]+2^{-2^{dL_n}}\sum_{i=1}^{2^{2^{L_n d}}}\mathbb{E}_{f_i}\left[1-\Psi_n\right]$$
$$\geq \int\left(\Psi_n(x_1,\dots,x_n)+1-\Psi_n(x_1,\dots,x_n)\right)$$
$$\left(\prod_{j=1}^n f_0(x_j)\wedge 2^{-2^{dL_n}}\sum_{i=1}^{2^{2^{L_n d}}}\prod_{j=1}^n f_i(x_j)\right)d\mu^{\otimes n}(x_1,\dots,x_n)$$
$$= 1-\frac{1}{2}\left\|P_0^{\otimes n}-Q^{\otimes n}\right\|_1$$
$$\geq 1-\frac{1}{2}\sqrt{\chi^2\left(Q^{\otimes n},P_0^{\otimes n}\right)}.$$

where $\chi^2(Q,P)=\int(dP/dQ-1)^2dQ$ if $P\ll Q$, $\chi^2(Q,P)=+\infty$ otherwise. Also, for any $1\leq\gamma,\kappa\leq 2^{2^{dL_n}}$, the orthonormality of the wavelet basis gives

$$\int\frac{dP_\gamma^{\otimes n}}{dP_0^{\otimes n}}\frac{dP_\kappa^{\otimes n}}{dP_0^{\otimes n}}dP_0^{\otimes n}$$
$$=\prod_{i=1}^n\int_{\mathbb{T}^d}\left[1+\epsilon 2^{-L_n(r+d/2)}\sum_k\omega_k^{(\gamma)}\psi_{L_n,k}(x_i)\right]\left[1+\epsilon 2^{-L_n(r+d/2)}\sum_k\omega_k^{(\kappa)}\psi_{L_n,k}(x_i)\right]dx_i$$
$$=\left(1+\epsilon^2 2^{-L_n(2r+d)}\sum_k\omega_k^{(\gamma)}\omega_k^{(\kappa)}\right)^n.$$

Then, for $\gamma_n = n\epsilon^2 2^{-L_n(2r+d)}\to 0$ and R_k, R_k' i.i.d. Rademacher random variables,

$$\chi^2\left(Q^{\otimes n},P_0^{\otimes n}\right)=2^{-2^{dL_n}}\sum_{\gamma,\kappa}\left(1+\epsilon^2 2^{-L_n(2r+d)}\langle\omega^{(\gamma)},\omega^{(\kappa)}\rangle\right)^n-1$$
$$\leq\mathbb{E}\left[\exp\left(n\epsilon^2 2^{-L_n(2r+d)}\sum_k R_kR_k'\right)\right]-1$$
$$=\mathbb{E}\left[\exp\left(n\epsilon^2 2^{-L_n(2r+d)}\sum_k R_k\right)\right]-1$$

$$= \cosh(\gamma_n)^{2^{L_n d}} - 1,$$

where we used that $1 + x \leq e^x$ for $x \in \mathbb{R}$ in the second line and that $R_k R'_k$ is distributed as R_k in the third. Using that $\cosh(z) = 1 + z^2/2 + \underset{|z|\to 0}{o}(z^2)$ and $1 + x \leq e^x$ once again, for any $\delta > 0$,

$$(\cosh(\gamma_n))^{2^{dL_n}} - 1 = \left(1 + \frac{\gamma_n^2}{2}(1 + o(1))\right)^{2^{dL_n}} - 1 \leq \exp\left(\gamma_n^2 2^{dL_n - 1}(1 + o(1))\right) - 1 \leq \delta^2$$

for n large enough, since $\gamma_n^2 2^{dL_n} = o(1)$. We have proven that, for any $\beta < 1$ and $\rho_n = o\left(\rho_n^*\right)$,

$$\liminf_n \inf_{\Psi_n} \left[\sup_{f \in \mathcal{F}(s)} \mathbb{E}_f\left[\Psi_n\right] + \sup_{f \in \tilde{\mathcal{F}}(r, \rho_n)} \mathbb{E}_f\left[1 - \Psi_n\right] \right] \geq \beta,$$

which concludes the proof. □

4.7 Proofs for Section 4.4

Proof of Proposition 8. As f and $\tilde{f}_n$ have the same total mass, we may without loss of generality take the supremum over functions $h \in \mathrm{Lip}_1(\mathbb{R}^d)$ for which $h(0) = 0$; observe that $x \mapsto \|x\|$ is an envelope for this function class. Since both f and $\tilde{f}_n$ have finite first moments (almost surely), the wavelet expansion of any h in this class converges in $L_1(f)$ and $L_1(\tilde{f}_n)$ and so

$$\int_{\mathbb{R}^d} h(f - \tilde{f}_n) = \sum_{k \in \mathbb{Z}^d} \langle h, \phi_k \rangle \langle f - \tilde{f}_n, \phi_k \rangle + \sum_{l \geq 0} \sum_{k \in \mathbb{Z}^d} \langle h, \psi_{lk} \rangle \langle f - \tilde{f}_n, \psi_{lk} \rangle.$$

As the father wavelets ϕ_k are compactly supported in some interval about k,

$$|\langle h, \phi_k \rangle| \lesssim |h(k)| \leq \|k\|$$

for some constant depending on the wavelet basis. Moreover, $h - K(h) = \sum_{l \geq 0} \sum_{k \in \mathbb{Z}^d} \langle h, \psi_{lk} \rangle \psi_{lk}$ is in a $B^1_{\infty\infty}$-ball of radius depending only on the wavelet basis, and so by (4.3),

$$\sup_{k \in \mathbb{Z}^d} |\langle h, \psi_{lk} \rangle| \lesssim 2^{-l\left(\frac{d}{2}+1\right)}.$$

Plugging these uniform estimates for the wavelet coefficients of h into the first equation gives the result. □

Proof of Theorem 20. When $d = 1$, the empirical measure achieves the stated rate ([60]). Thus we assume $d \geq 2$.

The estimator we use is

$$\hat{f}_n := \sum_{\|k\|_\infty \leq \kappa_{-1n}} \hat{f}_{-1k} \phi_k + \sum_{l \leq l_n(s)} \sum_{\|k\|_\infty \leq \kappa_{ln}} \hat{f}_{lk} \psi_{lk},$$

where $\hat{f}_{lk}$ are empirical wavelet coefficients and the cutoffs $\kappa_{ln}, l_n(s)$ are chosen such that

$$2^{l_n(s)} \simeq n^{\frac{1}{2s+d}}, \kappa_{-1n} = \kappa_{0n} \simeq (\log n)^\gamma, \kappa_{ln} = 2^l \kappa_{0n},$$

where γ is to be chosen below. We then use the decomposition in Proposition 8, which we further split to obtain six terms:

$$\begin{aligned} W_1(f,\hat{f}_n) \lesssim & \sum_{\|k\|_\infty \le \kappa_{-1n}} \|k\| |\hat{f}_{-1k} - f_{-1k}| + \sum_{\|k\|_\infty > \kappa_{-1n}} \|k\| |f_{-1k}| \\ & \ldots + \sum_{l<l_n(s)} 2^{-l\left(\frac{d}{2}+1\right)} \sum_{\|k\|_\infty \le \kappa_{ln}} |\hat{f}_{lk} - f_{lk}| + \sum_{l<l_n(s)} 2^{-l\left(\frac{d}{2}+1\right)} \sum_{\|k\|_\infty > \kappa_{ln}} |f_{lk}| \\ & \ldots + \sum_{l\ge l_n(s)} 2^{-l\left(\frac{d}{2}+1\right)} \sum_{\|k\|_\infty \le \kappa_{ln}} |f_{lk}| + \sum_{l\ge l_n(s)} 2^{-l\left(\frac{d}{2}+1\right)} \sum_{\|k\|_\infty > \kappa_{ln}} |f_{lk}| \\ & =: I + II + III + IV + V + VI. \end{aligned}$$

We first consider the bias terms II, IV, VI. For term II, we have that

$$\sum_{\|k\|_\infty > \kappa_{-1n}} \|k\| |f_{-1k}| \le \int_{\mathbb{R}^d} \sum_{\|k\|_\infty > \kappa_{-1n}} \|k\| |\phi_k(x)| f(x)\, \mathrm{d}x.$$

Since each ϕ_k is compactly supported in some interval about k, and $\sum_{k\in\mathbb{Z}^d} |\phi_k|$ is uniformly bounded on $\mathbb{R}^d$, we have that

$$\sum_{\|k\|_\infty > \kappa_{-1n}} \|k\| |\phi_k(x)| \lesssim \|x\|$$

for some constant depending on the wavelet basis. Moreover, the integrand is supported for all large enough n in $([-\kappa_{-1n}/2, \kappa_{-1n}/2]^d)^c =: D_n$. Thus, for n large enough,

$$II \lesssim \int_{D_n} \|x\| f(x)\, \mathrm{d}x \le \mathcal{E}_{\alpha,\beta}(f) \kappa_{-1n} \exp\left(-\beta\left(\frac{\kappa_{-1n}}{2}\right)^\alpha\right). \tag{4.43}$$

Since $\sum_{k\in\mathbb{Z}^d} |\psi_{lk}|$ is uniformly bounded by a constant depending on the wavelet basis times $2^{ld/2}$, we analogously have

$$\sum_{\|k\|_\infty > \kappa_{ln}} |f_{lk}| \lesssim 2^{ld/2} \mathcal{E}_{\alpha,\beta}(f) \exp\left(-\beta\left(\frac{\kappa_{0n}}{2}\right)^\alpha\right). \tag{4.44}$$

Thus

$$IV + VI \lesssim \mathcal{E}_{\alpha,\beta}(f) \exp\left(-\beta\left(\frac{\kappa_{0n}}{2}\right)^\alpha\right).$$

Choosing $\gamma > 0$ sufficiently large depending on α, β, these terms converge faster than $n^{-1/2}$.

Next, we deal with the final bias term V. By Cauchy-Schwarz and the fact that $\|f\|_{B^s_{2q}} \le B$,

$$\sum_{\|k\|_\infty \le \kappa_{ln}} |f_{lk}| \le \sqrt{\kappa_{ln}^d} \|f_{l\cdot}\|_2 \lesssim (\log n)^{\gamma d/2} 2^{l\left(\frac{d}{2}-s\right)},$$

and so

$$V \lesssim \sum_{l\ge l_n(s)} 2^{-l(s+1)} (\log n)^{\gamma d/2} \simeq (\log n)^{\gamma d/2} 2^{-l_n(s)(s+1)}$$

which is of the correct order by the definition of $l_n(s)$.

To bound the stochastic terms I and III, we use the expectation bound Lemma 35, whose proof generalises naturally to the case of $\mathbb{R}^d$. We have for all $l \ge -1$ such that $2^{ld} \le n$ and $k \in \mathbb{Z}^d$ that

$$E_f |\hat{f}_{lk} - f_{lk}| \lesssim n^{-1/2},$$

for some constant depending on M and the wavelet basis. So

$$E_f(I) \lesssim (\kappa_{-1n})^{d+1} n^{-1/2}$$

and

$$E_f(III) \lesssim (\log n)^{\gamma d} n^{-1/2} \sum_{l<l_n(s)} 2^{l\left(\frac{d}{2}-1\right)}.$$

When $d = 2$, the sum contributes an extra $\log n$ factor as in the statement. For $d \geq 3$, the final term of the sum dominates, and so

$$E_f(III) \lesssim (\log n)^{\gamma d/2} n^{-\frac{s+1}{2s+d}}$$

as stated. □

Proof of Theorem 21. Define the thresholds $\kappa_{-1n} = \kappa_{0n} \simeq (\log n)^\gamma$, $\kappa_{ln} = 2^l \kappa_{0n}$ for γ chosen as in Theorem 20. As before, let $l_{\max}$ be such that $2^{l_{\max}} \simeq (n/\log n)^{1/d}$; for $0 \leq l \leq l_{\max}$, define the thresholds τ_l via

$$\tau_l^2 = \tau^2 \kappa_{ln}^d \frac{\log n}{n},$$

where $\tau > 0$ is to be chosen below. For any sequence $(a_k)_{k\in\mathbb{Z}^d}$, set $\|a_\cdot\|_{2,\kappa_{ln}} := \left(\sum_{\|k\|_\infty \leq \kappa_{ln}} a_k^2\right)^{1/2}$. The thresholded estimator is then defined as

$$\hat{f}_n = \sum_{\|k\|_\infty \leq \kappa_{-1n}} \hat{f}_{-1k} \phi_k + \sum_{l=0}^{l_{\max}} \mathbb{1}_{\left\{\|\hat{f}_{l\cdot}\|_{2,\kappa_{ln}} > \tau_l\right\}} \sum_{\|k\|_\infty \leq \kappa_{ln}} \hat{f}_{lk} \psi_{lk}. \tag{4.45}$$

We perform a decomposition of the risk similar to that in the previous proof:

$$\begin{aligned}
W_1(f, \hat{f}_n) \lesssim & \sum_{\|k\|_\infty \leq \kappa_{-1n}} \|k\| |\hat{f}_{-1k} - f_{-1k}| + \sum_{\|k\|_\infty > \kappa_{-1n}} \|k\| |f_{-1k}| \\
& + \sum_{l \leq l_{\max}} 2^{-l\left(\frac{d}{2}+1\right)} \sum_{\|k\|_\infty \leq \kappa_{ln}} \left| \mathbb{1}_{\left\{\|\hat{f}_{l\cdot}\|_{2,\kappa_{ln}} > \tau_l\right\}} \hat{f}_{lk} - f_{lk} \right| \\
& + \sum_{l \leq l_{\max}} 2^{-l\left(\frac{d}{2}+1\right)} \sum_{\|k\|_\infty > \kappa_{ln}} |f_{lk}| \\
& + \sum_{l > l_{\max}} 2^{-l\left(\frac{d}{2}+1\right)} \sum_{\|k\|_\infty \leq \kappa_{ln}} |f_{lk}| + \sum_{l > l_{\max}} 2^{-l\left(\frac{d}{2}+1\right)} \sum_{\|k\|_\infty > \kappa_{ln}} |f_{lk}| \\
=: & \; I + II + III + IV + V + VI.
\end{aligned}$$

We treat terms I, II, IV and VI identically to before. Term V is also dealt with as in the previous proof, noting that for all n sufficiently large, $2^{l_{\max}} > n^{1/(2s+d)}$. It remains to deal with term III; by Cauchy-Schwarz and the definition of κ_{ln}, we have that

$$III \lesssim (\log n)^{\frac{\gamma d}{2}} \sum_{l=0}^{l_{\max}} 2^{-l} \left\| \mathbb{1}_{\left\{\|\hat{f}_{l\cdot}\|_{2,\kappa_{ln}} > \tau_l\right\}} \hat{f}_{l\cdot} - f_{l\cdot} \right\|_{2,\kappa_{ln}},$$

where the constant depends on d. By splitting this sum into two parts at $l_n(s)$,

$$2^{l_n(s)} \simeq B^{1/s} (n/\log n)^{1/(2s+d)},$$

one can bound it exactly as in the proof of Theorem 17 □

Proof of Theorem 22. We first establish coverage. Define the thresholds κ_{ln} as in the previous proof. Given $f \in \mathcal{G}(s)$, as in the proof of Theorems 17 and 21, on an event of probability tending to 1, for all l such that $l_n(r) \leq l \leq l_{\max}$, $\langle \hat{f}_n, \psi_{l\cdot} \rangle \equiv 0$. Note that $l_{\max} > j_n > l_n(s) > l_n(r)$. So on this event, by Cauchy-Schwarz,

$$\begin{aligned}\left(\sum_{l \geq j_n} 2^{-l\left(\frac{d}{2}+1\right)} \sum_{\|k\|_\infty \leq \kappa_{ln}} |\langle f - \hat{f}_n, \psi_{lk} \rangle| \right)^2 &\lesssim (\log n)^{\gamma d} \left(\sum_{l \geq j_n} 2^{-l} \|\langle f, \psi_{l\cdot} \rangle\|_2 \right)^2 \\ &\lesssim (\log n)^{\gamma d} B 2^{-2 j_n (r+1)} \\ &\leq G_{j_n}\end{aligned}$$

for all n sufficiently large, i.e. this quantity is $O_P(G_{j_n})$. The other term in $\tilde{W}^{(n)}(f, \hat{f}_n)^2$ is precisely $E_f^{(2)} V_{j_n}$; by Chebyshev's inequality we obtain condition (4.23).

It remains to confirm the diameter conditions (4.24) and (4.25) with the rates $R_n(r), R_n(s)$ as given in the statement of the result. As the remainder term $\sqrt{r_n}$ converges up to a logarithmic factor at the rate $n^{-1/2}$, it is dominated by $\tilde{W}^{(n)}$ for diameter considerations. As observed previously, we may instead prove the diameter conditions for the $\tilde{W}^{(n)}$ distance with the augmented rates

$$\bar{R}_n(r) = (\log n)^{\gamma d/2} \left(\frac{n}{\log n} \right)^{-\frac{r+1}{2r+d}}, \quad \bar{R}_n(s) = (\log n)^{\gamma d/2} \left(\frac{n}{\log n} \right)^{-\frac{s+1}{2s+d}}.$$

By the same argument as in the proof of Theorem 18, the $\tilde{W}^{(n)}$-diameter of C_n is bounded by a constant multiple of

$$(\log n)^{\gamma d/4 + 1/2} n^{-1/2} \left(\sum_{l<j} 2^{l(d-4)} \right)^{1/4} + \sqrt{V_{j_n}} + \sqrt{G_{j_n}} + n^{-1/2}.$$

The final term is dominated by the first, and (using the condition on s when $d > 4$) $\sqrt{G_{j_n}} = O(\bar{R}_n(s)) = o(\bar{R}_n(r))$. One checks the first term is of the correct order as in Theorem 18. Finally, since $\mathrm{Var}_f^{(2)}(V_{j_n}) \to 0$ (one shows that $\tilde{W}^{(n,j_n)}(f, \hat{f}_n) \to 0$ analogously to the proof of Theorem 17), we have that

$$V_{j_n} = O_{P_f}\left(E_f V_{j_n}\right);$$

as in the proof of Theorem 21, this expectation is of order $\bar{R}_n(r)$ or $\bar{R}_n(s)$ when f belongs to $\mathcal{G}(r)$ or $\mathcal{G}(s)$ respectively. □

Proof of Theorem 23. For some $\alpha' > \alpha$, $D > 0$ and $\alpha(x) = \alpha' e^{-1/\left(\|x\|_2 - D\right)} \mathbb{1}_{B(0,D)^c}(x)$, the density f defined by

$$f(x) \propto e^{-\beta \|x\|_2^{\alpha(x)}}$$

is such that $\mathbb{E}_f\left[e^{\beta \|X\|^\alpha}\right] < +\infty$. Then, for $\sigma > 0$, if $X \sim P_f$, σX has density $g : x \mapsto \sigma^{-d} f(\sigma^{-1} x)$ satisfying

$$\mathbb{E}_g\left[e^{\beta \|X\|^\alpha}\right] = \mathbb{E}_f\left[e^{\sigma^\alpha \beta \|X\|^\alpha}\right] < +\infty.$$

Then, we verify that $f \in H_2^m(\mathbb{R}^d) \subset B_{2\infty}^m(\mathbb{R}^d) \subset B_{2q}^s(\mathbb{R}^d)$, for any $m \in N$ and $s < m$. Also, $\|g\|_p = \sigma^{-d(1-1/p)} \|f\|_p$ and, the moduli of continuity of g satisifies, for $t > 0$ and an integer $r > s$,

$$\begin{aligned}\omega_r(g,t,2) &:= \sup_{0\le\|h\|\le t} \left\| \sum_{k=0}^r \binom{r}{k} (-1)^{r-k} \sigma^{-d} f(\sigma^{-1}\cdot + k\sigma^{-1}h) \right\|_2 \\ &= \sigma^{-d} \sup_{0\le\|h\|\le \sigma^{-1}t} \left\| \sum_{k=0}^r \binom{r}{k} (-1)^{r-k} \sigma^{d} f(\sigma^{-1}\cdot + kh) \right\|_2 \\ &= \sigma^{-d/2} \omega_r(f, \sigma^{-1}t, 2).\end{aligned}$$

Therefore, with $|f|_{B_{pq}^s} := \left[\int_0^\infty \left|\frac{\omega_r(f,t,p)}{t^s}\right|^q \frac{dt}{t}\right]^{1/q}$, we have

$$\|g\|_{B_{pq}^s} = \|g\|_p + |g|_{B_{pq}^s} = \sigma^{-d(1-1/p)} \|f\|_p + \sigma^{-d(1-1/p)-s} |f|_{B_{pq}^s}, \tag{4.46}$$

so that $\|g\|_{B_{2q}^s} \le B$ for σ large enough. Also, since $f \in L_\infty(\mathbb{R}^d)$, $g \le M$ for σ large enough. So, for some large L, $g \in \mathcal{G}(s)$.

For some sequence $L_n \to \infty$, and any $\omega \in \{-1;1\}^{\mathbb{Z}\cap\left[0,2^{L_n}\right)^d \times \mathcal{I}}$, we define for some $\epsilon > 0$,

$$f_\omega^n = g + \epsilon 2^{-L_n(r+d/2)} \sum_{k\in\mathbb{Z}\cap[0,2^{L_n})^d, \iota\in\mathcal{I}} \omega_{k,\iota} \Psi_{L_n k}^\iota.$$

Assuming that the scaling and mother wavelets functions are compactly supported (as assumed in Appendix 4.5), the $\Psi_{L_n k}^\iota$, for $k \in \mathbb{Z} \cap \left[0, 2^{L_n}\right)^d$, $\iota \in \mathcal{I}$, are supported on a compact set K independent of n. Then, since

$$\begin{aligned}\|f_\omega^n\|_{B_{2q}^r} &\le \|g\|_{B_{2q}^r} + \epsilon 2^{-L_n(r+d/2)} \left\| \sum_{k\in\mathbb{Z}\cap[0,2^{L_n})^d, \iota\in\mathcal{I}} \omega_{k,\iota} \Psi_{L_n k}^\iota \right\|_{B_{2q}^r} \\ &\le \|g\|_{B_{2q}^r} + C\epsilon,\end{aligned}$$

for some $C > 0$ depending on d only, reasoning as for (4.46), and since $B_{2q}^s \subset B_{2q}^r$, f_ω^n is the $\|\cdot\|_{B_{2q}^r}$-Besov ball of radius B for ϵ small enough and σ large enough. Then, by assumption, $\int_{\mathbb{R}^d} f_\omega^n(t)dt = 1$ and, since

$$\left\| 2^{-L_n(r+d/2)} \sum_{k\in\mathbb{R}^d, \iota\in\mathcal{I}} \left|\Psi_{L_n k}^\iota\right| \right\|_\infty \lesssim 2^{-rL_n},$$

$0 < f_\omega^n \le M$ for n, σ large enough (or ϵ small enough if $r = 0$). Indeed, g is lower bounded by a some positive constant on K, So, f_ω^n actually is a density function.

For these to belong to the alternative hypothesis, it remains to check that these are well separated from the null hypothesis. For any $h \in \mathcal{G}(s)$, the reversed triangular inequality gives

$$\begin{aligned}W_1(f_\omega^n, h) &\gtrsim \|f_\omega^n - h\|_{B_{1\infty}^{-1}} \\ &\gtrsim 2^{-L_n(d/2+1)} \sum_{k\in\mathbb{Z}\cap[0,2^{L_n})^d, \iota\in\mathcal{I}} \left|\langle f_\omega^n - h, \Psi_{L_n k}^\iota\rangle\right|\end{aligned}$$

$$
\begin{aligned}
&\geq 2^{-L_n(d/2+1)} \left| \sum_{k,\iota} |\langle f_\omega^n - g, \Psi_{Lnk}^\iota \rangle| - \sum_{k,\iota} |\langle h - g, \Psi_{Lnk}^\iota \rangle| \right| \\
&= 2^{-L_n(d/2+1)} \left[C 2^{-L_n(r-d/2)} - C' 2^{-L_n(s-d/2)} \right] \\
&\gtrsim 2^{-L_n(1+r)},
\end{aligned}
$$

for constants independent of n. Above, we used that $s > r$ and that, for any $s > 0$, $\mathcal{G}(s) \subset \left\{ f : \|f\|_{B_{1q}^s} \leq B' \right\}$ for some $B' > 0$ according to Lemma 4.20.

The last inequality holds for n large enough. Therefore, if L_n^* is such that $2^{-L_n^*(1+r)} \asymp \xi_n$, it is possible to take $L_n > L_n^*$ such that $\rho_n \leq C' 2^{-L_n(1+r)} = o(\xi_n)$, so that, for any ω, $f_\omega^n \in \tilde{\mathcal{G}}(s, \rho_n)$.

For $N_n = 2^{2^{dL_n}\left(2^d-1\right)}$, let's index $\omega \in \{-1;1\}^{\mathbb{Z} \cap \left[0, 2^{L_n}\right)^d \times \mathcal{I}} = \left\{ w^{(m)} : \ m = 1, \dots, N_n \right\}$ and denote $P_m = P_{f_{\omega^{(m)}}}$. Then,

$$
\liminf_n \inf_{\Psi_n} \left[\sup_{f \in H_0} \mathbb{E}_f \left[\Psi_n \right] + \sup_{f \in H_1(r_n)} \mathbb{E}_f \left[1 - \Psi_n \right] \right] \geq 1 - \frac{1}{2} \sqrt{\chi^2 \left(Q^{\otimes n}, P_0^{\otimes n} \right)},
$$

where $Q = N_n^{-1} \sum_{m=1}^{N_n} P_m$ and P_0 has density $g \in H_0$. Then, for any $1 \leq m, q \leq N_n$, one has by properties of the wavelet basis, denoting $\nu_m = f_{\omega^{(m)}} - g$,

$$
\begin{aligned}
&\int \frac{dP_m^{\otimes n}}{dP_0^{\otimes n}} \frac{dP_q^{\otimes n}}{dP_0^{\otimes n}} dP_0^{\otimes n} \\
&= \prod_{i=1}^n \int_{[0,1]^d} \left[g(x_i) + \epsilon 2^{-L_n(r+d/2)} \sum_{k,\iota} \omega_{k,\iota}^{(m)} \Psi_{L_n k}^\iota(x_i) \right] \\
&\qquad \left[g(x_i) + \epsilon 2^{-L_n(r+d/2)} \sum_{k,\iota} \omega_{k,\iota}^{(q)} \Psi_{L_n k}^\iota(x_i) \right] g^{-1}(x_i) dx_i \\
&= \left(1 + \int_{\mathbb{R}^d} \frac{\nu_m(x) \nu_q(x)}{g(x)} dx \right)^n.
\end{aligned}
$$

For σ large enough, g is constant on the compact support of ν_m and ν_q, equal to $g(0)$. Hence, following the same arguments as above,

$$
\chi^2 \left(Q^{\otimes n}, P_0^{\otimes n} \right) = (\cosh \gamma_n)^{2^{dL_n}\left(2^d-1\right)} - 1,
$$

where $\gamma_n = n \epsilon^2 g(0)^{-1} 2^{-L_n(2r+d)}$, and for any $\delta > 0$, $\chi^2 \left(Q^{\otimes n}, P_0^{\otimes n} \right) \leq \delta^2$ for n large enough. This concludes the proof. □

Lemma 38. *Let $B \geq 1, M > 0, \alpha > 0, \beta > 0, L > 0, 1 \leq q \leq \infty$, and $s \geq 0$. Then, there exists a constant B', depending on the class parameters, the wavelet basis and the dimension d, such that*

$$
\mathcal{G}_{s,2,q}(B, M; \alpha, \beta, L) \subset \mathcal{G}_{s,1,q}(B', M; \alpha, \beta, L).
$$

Proof. Let $f \in \mathcal{G}_{s,2,q}(B, M; \alpha, \beta, L)$. All we have to prove is that

$$
\|f\|_{B_{1q}^s} = \|\langle f, \phi_\cdot \rangle\|_1 + \left(\sum_{l \geq 0} \left[2^{l(s-d/2)} \|\langle f, \psi_{l\cdot} \rangle\|_1 \right]^q \right)^{1/q} \leq B',
$$

for some B' as in the lemma. Let $\kappa > 0$. Then,

$$\|\langle f, \phi_\cdot \rangle\|_1 = \sum_{\|k\|_\infty \leq \kappa} |\langle f, \phi_k \rangle| + \sum_{\|k\|_\infty > \kappa} |\langle f, \phi_k \rangle| .$$

For the second term, the same arguments as the one used to obtain (4.43) give that it is bounded by $\mathcal{E}_{\alpha,\beta}(f) \exp\left(-\beta \left(\frac{\kappa}{2}\right)^\alpha\right)$, up to a constant depending on the wavelet basis. The first term is controlled via the Cauchy-Schwarz inequality

$$\sum_{\|k\|_\infty \leq \kappa} |\langle f, \phi_k \rangle| \lesssim (2\kappa+1)^d \left(\sum_{\|k\|_\infty \leq \kappa} |\langle f, \phi_k \rangle|^2 \right)^{1/2} \leq (2\kappa+1)^d \|\langle f, \phi_\cdot \rangle\|_2 ,$$

for a constant depending on d only.

Next consider, for $l \geq 0$, $\|\langle f, \psi_{l\cdot} \rangle\|_1$. As before, letting $\kappa_l = 2^{l/2}$, we have

$$\|\langle f, \psi_{l\cdot} \rangle\|_1 = \sum_{\|k\|_\infty \leq \kappa_l} |\langle f, \psi_{lk} \rangle| + \sum_{\|k\|_\infty > \kappa_l} |\langle f, \psi_{lk} \rangle| .$$

Arguing as with (4.44), the second term is bounded by $2^{ld/2} \mathcal{E}_{\alpha,\beta}(f) \exp\left(-\beta \left(\frac{\kappa_l}{2}\right)^\alpha\right)$, up to a constant depending on the wavelet basis. The first term is controlled as above. Then, using the l^q triangular inequality,

$$\begin{aligned}
&\left(\sum_{l \geq 0} \left[2^{l(s-d/2)} \|\langle f, \psi_{l\cdot} \rangle\|_1 \right]^q \right)^{1/q} \\
&\lesssim \left(\sum_{l \geq 0} 2^{ql(s-d/2)} \left[2^{ld/2} \mathcal{E}_{\alpha,\beta}(f) \exp\left(-\beta \left(\frac{\kappa_l}{2}\right)^\alpha\right) + (2\kappa_l+1)^d \|\langle f, \psi_{l\cdot} \rangle\|_2 \right]^q \right)^{1/q} \\
&\lesssim \left(\sum_{l \geq 0} \left[2^{ls} \mathcal{E}_{\alpha,\beta}(f) \exp\left(-2^{-\alpha} \beta 2^{l\alpha/2}\right) \right]^q \right)^{1/q} + \left(\sum_{l \geq 0} \left[2^{ls} \|\langle f, \psi_{l\cdot} \rangle\|_2 \right]^q \right)^{1/q},
\end{aligned}$$

for a constants depending on the wavelet basis and d. The first term is upper bounded by

$$\mathcal{E}_{\alpha,\beta}(f) \left(\sum_{l \geq 0} 2^{qls} \exp\left(-q 2^{-\alpha} \beta 2^{l\alpha/2}\right) \right)^{1/q} \lesssim \mathcal{E}_{\alpha,\beta}(f),$$

as the series converges.

In the end, following our assumptions on $\|f\|_{B^s_{2q}}$,

$$\begin{aligned}
\|f\|_{B^s_{1q}} &\lesssim (2\kappa+1)^d \|\langle f, \phi_\cdot \rangle\|_2 + \left(\sum_{l \geq 0} \left[2^{ls} \|\langle f, \psi_{l\cdot} \rangle\|_2 \right]^q \right)^{1/q} + \mathcal{E}_{\alpha,\beta}(f) \exp\left(-\beta \left(\frac{\kappa}{2}\right)^\alpha\right) \\
&\quad + \mathcal{E}_{\alpha,\beta}(f) \\
&\lesssim B + \mathcal{E}_{\alpha,\beta}(f) \leq B + L,
\end{aligned}$$

where the constants depend on the wavelet basis, d, the arbitrary $\kappa > 0$ we took, s, α, β and q.

□

CHAPTER 5

Deep Horseshoe Gaussian Processes

This work is concerned with the study of theoretical properties of deep Gaussian processes, which have recently been proposed as natural objects to fit, similarly to deep neural networks, possibly complex features present in modern data samples, such as compositional structures. Adopting a Bayesian nonparametric approach, it is natural to use deep Gaussian processes as prior distributions, and to use the corresponding posterior distributions for statistical inference. We introduce the deep Horseshoe Gaussian process Deep–HGP, a new prior based on deep Gaussian processes with squared-exponential kernel, that in particular enables data-driven choices of the key lengthscale parameters. For nonparametric regression with random design, we show that the associated tempered posterior distributions recovers the unknown true regression curve optimally in terms of quadratic loss, up to a logarithmic factor. At the same time, Deep–HGP are conceptually quite simple to construct. One main idea is that the horseshoe prior enables *simultaneous* adaptation to both smoothness *and* structure.

Table of Contents

5.1 Introduction

5.1.1 Gaussian processes

Gaussian processes (henceforth GPs) are among the most used machine learning methods, with applications ranging from inference in regression models to classification, see e.g. [138] for an overview. Due to their flexibility, in recent years GPs have been used as tools for geometric inference and deep learning.

A particularly natural field of application where there now exists at least partial theory to explain and validate practical successes of GPs is that of *Bayesian nonparametrics*, where the posterior distribution can also be used for the practically essential task of *uncertainty quantification*. In regression settings, it is particularly natural to use a GP as a prior distribution on the unknown regression function. The corresponding posterior distributions can often be efficiently implemented (in particular for regression with Gaussian noise, for which GPs form a conjugate class of prior distributions) and come with theoretical convergence guarantees: the works [163, 160, 27] indeed show that the convergence rate in typical metrics of the posterior distribution is completely determined (both upper and lower bounds) by the behaviour of its concentration function. Shortly thereafter, van der Vaart and van Zanten also showed that statistical *adaptation to smoothness* was possible with GPs with optimal minimax contraction rates by simply drawing at random its scaling parameter [165] in fixed design regression; see [161, 130] for extensions to random design regression and [155] to inverse problems.

Let us mention a few applications of posterior distributions arising from GPs that are related to the setting considered below.

GPs flexibility: geometric setting. In modern statistical settings, it is frequent that data naturally sit on a geometric object such as a compact manifold (e.g. a sphere, swissroll etc.). It is tempting to use GPs in this setting as well, although some care is needed in their construction. For instance, the celebrated GP with squared–exponential kernel (thereafter SqExp) has no immediate analog in a manifold setting, as replacing the euclidian metric in the exponential defining SqExp by the geodesic distance does not form a covariance kernel. This can be remediated by using a kernel coming from heat equation solutions on the manifold [30], and this kernel can be shown to be a natural geometric analog of SqExp. Alternatively, one may put a prior directly on the ambient space equipped with the standard euclidean metric: the authors in [174] obtain a posterior rate that under some (smoothness) conditions adapts to the unknown dimension of the manifold with a rescaled SqExp exponential GP, when the loss function is the quadratic loss but restricted to sit on the manifold.

GPs and adaptation to anisotropy. By drawing independent lengthscale parameters along different dimensions, [9] show that posteriors arising from SqExp GPs contract at near-optimal minimax anisotropic rates. A related problem is that of variable selection in (possibly high-dimensional regression). The unknown regression function may indeed depend only on a few coordinates (although these are not known in advance).

GPs and variable selection. By considering variable selection type priors and then drawing lengthscale parameters of SqExp GPs, [176] and [84] provide theory for this setting and respectively investigate optimal rates and variable selection properties for the corresponding posterior distributions.

5.1.2 Deep Gaussian processes

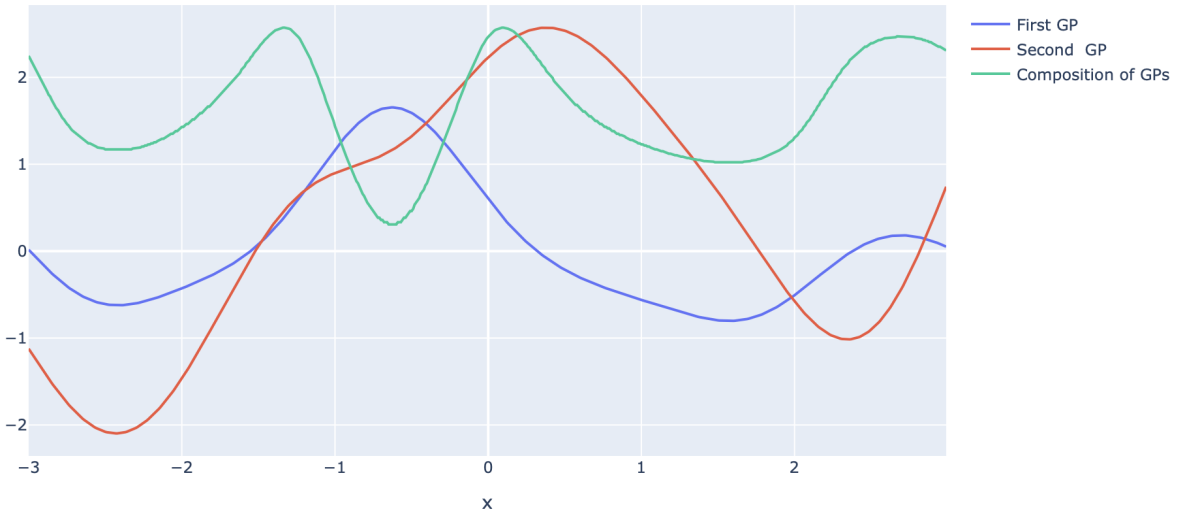
Recent years have seen a number of remarkable applications of so–called *deep learning* methods, where 'deep' typically refers to a certain (often compositional) structure in terms of a number of layers (for instance, deep neural networks).

Deep Gaussian processes correspond to iterated compositions of Gaussian processes and broadly speaking can be seen as (one possible) Bayesian analogue of deep neural networks. There is a lot of recent activity for providing efficient sampling methods for deep GPs. Yet, theory is just starting to emerge.

Focusing on compositions of constrained GPs (with bounded sample paths and derivatives), the recent paper [6] introduced a new concentration function for deep GPs, extending the GP contraction rate theory. The contraction rates they obtain via this method are shown to be minimax optimal for a variety of kernel choices and models.

Another recent work by Finocchio and Schmidt–Hieber [59] shows that using a model selection prior to select variables, and conditioning individual Gaussian process sample paths to verify certain smoothness constraints, the induced posterior distributions contract nearly optimally in quadratic loss for compositional structures. This work follows their footsteps and aims at answering the following question. Is it possible to simplify the prior construction to make it close to what is actually used in practice, while at the same time keeping similar optimal theoretical guarantees? While the prior proposed in [59] is completely natural and 'canonical' from the theoretical perspective, both the conditioning step (to match smoothness constraints) and the model selection prior (for which the posterior on submodels is often expensive to compute) make posterior sampling more involved in view of practical implementation. One main aim here is to try to simplify the construction of the prior as much as possible while keeping optimality, and thereby come closer to the practically used deep GPs, for which lengthscale parameters are often kept free and then adjusted empirically [44].

Figure 5.1: Composition of two Gaussian processes with SqExp covariance kernel $K(s,t) = e^{-(s-t)^2}$.



5.2 The setting and a novel prior

Consider the nonparametric regression problem with random design, where one observes $(X, Y) := (X_i, Y_i)_{1 \leq i \leq n}$, with $X_1, \ldots, X_n$ independent identically distributed design points sampled from a probability measure μ on $[-1, 1]^d$ and

$$Y_i = f_0(X_i) + \varepsilon_i, \tag{5.1}$$

for $f_0 : [-1, 1]^d \to \mathbb{R}$ an unknown regression function and ε_i independent $\mathcal{N}(0, \sigma_0^2)$ errors, with σ_0 assumed known for simplicity. We consider estimation of f_0 with respect to the integrated quadratic loss

$$\|f_0 - f\|^2_{L^2(\mu)} = \int (f_0 - f)^2 d\mu.$$

For a given regression function f, let P_f denote the distribution of one observation (X_i, Y_i) under model (5.1), which has density

$$p_f(x, y) = \left(2\pi\sigma_0^2\right)^{-1/2} e^{-\frac{(y - f(x))^2}{2\sigma_0^2}}$$

with respect to $\mu \otimes \lambda$, for λ the Lebesgue measure on $\mathbb{R}$.

For a real $\beta > 0$ and r an integer, let $\mathcal{C}^\beta[-1, 1]^r$ denote the classical Hölder space equipped with the norm $\|\cdot\|_{\beta,\infty}$. It consists of functions $f : [-1, 1]^r \to \mathbb{R}$ whose norm defined as

$$\|f\|_{\beta,\infty} = 2r \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|<\lfloor\beta\rfloor} \|\partial^{\boldsymbol{\alpha}} f\|_\infty + 2^{\beta - \lfloor\beta\rfloor} \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor\beta\rfloor} \sup_{\boldsymbol{x},\boldsymbol{y}\in[-1,1]^r,\ \boldsymbol{x}\neq\boldsymbol{y}} \frac{|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{y})|}{|\boldsymbol{x} - \boldsymbol{y}|_\infty^{\beta - \lfloor\beta\rfloor}}$$

is finite, with the multi-index notation $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r) \in \mathbb{N}^r$, $|\boldsymbol{\alpha}| := |\boldsymbol{\alpha}|_1$ and $\partial^{\boldsymbol{\alpha}} = \partial^{\boldsymbol{\alpha}_1} \ldots \partial^{\boldsymbol{\alpha}_r}$. We note, for $\beta' \leq \beta$, that $\|f\|_{\beta',\infty} \leq \|f\|_{\beta,\infty}$ according to Lemma 10 of [147].

5.2.1 Structural assumptions for multivariate regression

In order to assess the performance of machine learning methods, a popular benchmark is the regression setting (5.1) equipped with some ' structural' assumptions. In the unconstrained case where only smoothness is assumed on f_0, rates for β–Hölder smooth functions are typically of the form $n^{-\beta/(2\beta+d)}$, and so are prone to the curse of dimensionality (the rate becomes extremely slow for large d). A common approach is to assume that the multivariate regression function f_0 admits a certain *unknown* 'structure', possibly of 'dimension d^*' and possibly much smaller than d. For instance, in the simplest setting considered below, f_0 may only depend on a small but unknown number of coordinates.The goal is then to find algorithms that are able to achieve optimal risk bounds that adapt to the unknown underlying structure, and that therefore scale with d^* instead of d.

A first basic setting: shallow variable selection. Let us first consider the simple setting where $f_0 : [-1, 1]^d \to \mathbb{R}$ only depends on $d^* \leq d$ variables, that is

$$f_0(x_1, \ldots, x_d) = g(x_{i_1}, \ldots, x_{i_{d^*}}),$$

for some $g \in \mathcal{C}^\beta[-1,1]^{d^*}$, for some $\beta > 0$. The subset of indices $i_1, \dots, i_{d^*}$ is unknown to the statistician and the target convergence rate in quadratic loss is $n^{-\beta/(2\beta+d^*)}$, which can be much smaller than $n^{-\beta/(2\beta+d)}$.

Compositional structure. Following [147], suppose that f can be written as a composition

$$f = h_q \circ \cdots \circ h_0,$$

with $h_i : [-1,1]^{d_i} \to [-1,1]^{d_{i+1}}$, for (d_i) a sequence of integers such that $d_0 = d$ and $d_{q+1} = 1$. Since h_i takes values in $\mathbb{R}^{d_{i+1}}$, one may write $h_i = (h_{ij})$, where h_{ij} for $j = 1, \dots, d_{i+1}$ are its 1–dimensional components. Let us further assume that these only depend on a subset $\mathcal{S}_i$ of at most $t_i \le d_i$ variables, and h_{ij} restricted to $\mathcal{S}_i$ belongs to $\mathcal{C}^{\beta_i}[-1,1]^{t_i}$ as defined above and suppose, for any i, j,

$$\|h_{ij}\|_{\beta_i,\infty} \le K,$$

for some unknown $K > 0$.

The compositional class $\mathcal{F}(\lambda, \beta, K)$. In the setting of the previous paragraph, we denote by $\lambda = (q, d_1, \dots, d_q, t_0, \dots, t_q)$ the parameters describing this compositional structure and, for $\beta = (\beta_0, \dots, \beta_q)$, we let $\mathcal{F}(\lambda, \beta, K)$ denote the set of densities verifying the above conditions.

Minimax optimal rate. The minimax rate of estimation in quadratic loss over this class

$$(\tau_n^*)^2 = \inf_T \sup_{f \in \mathcal{F}(\lambda,\beta,K)} E_f \|T - f\|_2^2,$$

for T an arbitrary estimator of f, is, up to logarithmic factors, see [147],

$$\tau_n^* \asymp \max_{i=0,\dots,q} n^{-\frac{\beta_i \alpha_i}{2\beta_i\alpha_i + t_i}}, \qquad \text{where } \alpha_i := \prod_{l=i+1}^{q} (\beta_l \wedge 1).$$

5.2.2 Key ingredients

Posterior distributions: frequentist analysis. Given a prior distribution Π on regression functions, the posterior distribution is $\Pi[\cdot \,|\, X, Y]$ is given by Bayes' formula: this is the next display for $\rho = 1$. More generally, one may set, for any $\rho \in (0,1)$ and a measurable set B,

$$\Pi^\rho[B \,|\, X, Y] = \frac{\int_B \prod_{1 \le i \le n} p_f(X_i, Y_i)^\rho d\Pi(f)}{\int \prod_{1 \le i \le n} p_f(X_i, Y_i)^\rho d\Pi(f)}.$$

When $\rho = 1$ this is the usual posterior. If $0 < \rho < 1$, this quantity is called *ρ–posterior* (or tempered posterior). Its use is very much widespread in machine learning, in particular in PAC–Bayesian settings. We will use the tempered posterior in our main result, and discuss in more details its links with the case $\rho = 1$ in Section 5.4, but already note that computationally ρ–posteriors do not bring difficulties, and theory is similar (even simpler).

Gaussian process (ρ–) posteriors: theory. For any Gaussian process W on the Banach space of continuous functions equipped with the $\|\cdot\|_\infty$ norm, the probability measure of any ball $\{f : \|f - g\|_\infty < \varepsilon\}$ is lower bounded by a quantity depending on the mass of the centered ball of radius ε and on how well g can be approximated by elements of the RKHS $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$

corresponding to the covariance kernel of the process. More precisely, according to Proposition 11.19 of [65], we have

$$P\left[\|W-g\|_\infty < \varepsilon\right] \geq e^{-\varphi_g(\varepsilon)},$$
$$\varphi_g(\varepsilon) = \inf_{h\in\mathbb{H}:\ \|h-g\|_\infty \leq \varepsilon} \|h\|_{\mathbb{H}}^2 - \log P\left[\|W\| < \varepsilon\right]. \tag{5.2}$$

In nonparametric regression with fixed design, [165] proved that adaptive posterior contraction rates are achievable for stationary Gaussian process priors, with a dilatation parameter of the sample paths distributed as a Gamma variable. As a particular case, consider the *squared exponential* process SqExp defined as the zero-mean Gaussian process with covariance kernel $K(s,t) = \exp(-\|t-s\|^2)$ (and $\|\cdot\|$ the euclidean norm) on $[-1,1]^d$. Next, for $k,\theta>0$, one sets

$$A^d \sim \text{Gamma}\,(k,\theta)$$
$$f \;\Big|\; A \sim \left\{W_{At}:\ t\in[0,1]^d\right\}.$$

This construction induces a prior on the Banach space of continuous functions for which the posterior concentrates as in the empirical L_2-norm at rate $\varepsilon_n \asymp n^{-\beta/(2\beta+d)}$ (up to a log factor) whenever f_0 is β-Hölder regularity, $\beta>0$.

Although this rate coincides with the minimax estimation rate over a ball in $\mathcal{C}^\beta[0,1]^d$, it may be quite slow for large d. However, if the fixed design is located on a d^*-dimensional Riemannian manifold, $d^*<d$, of the ambient space $[0,1]^d$, we expect the faster rate $n^{-\beta/(2\beta+d^*)}$ to be attainable. The work [174] achieves it with a dilated Gaussian process as well, the dilatation factor A being distributed as $A^{d^*} \sim \text{Gamma}\,(k,\theta)$. A first remark on this method is that it needs an estimate of d^* to be applied. Secondly, the posterior contraction rates are obtained for local distances *on the manifold* (such as the empirical L_2-norm) only, but not on the ambient space.

When the regression function f_0 depends on a small number of variables d^* only, a simplified version of the add-GP prior from [176] gives optimal posterior contraction rates $n^{-\beta/(2\beta+d^*)}$ without the need to estimate d^*. This is achieved by the introduction of an additional layer in the prior, drawing via Bernoulli random variables in which direction the Gaussian sample paths have to be dilated (the sample paths being constant in the other directions). From a practical point-of-view, this 'hard' selection of variables adds a combinatorial complexity to posterior sampling.

Below, we introduce the Horseshoe Gaussian process prior to answer the question of variable selection and posterior contraction rates for the most natural global L^2 loss (in contrast to a loss e.g. restricted only on active directions) via a soft selection of dilated variables.

5.2.3 Deep Horseshoe Gaussian Process prior

We introduce a Gaussian process prior with independent lengthscales distributed following a half-horseshoe distribution. This distribution possesses two interesting properties for our goals. Its density has a pole at 0, which allows to 'select' irrelevant dimensions, drawing small lengthscales with high probability. It also has heavy Cauchy-like tails, so that it performs an adequate scaling on the ambient dimensions with sufficiently large probability. This distribution is then a favorable one for high-dimensional settings where d can be much larger than n. Although we do not tackle this setting in this chapter, we keep this for future works and we

explicit the dependence in the dimension in the proofs of our main results in preparation for these applications.

The single-layer case. Define the *Horseshoe Gaussian Process* prior HGP as the prior Π on regression functions f of the form

$$\begin{aligned} A_j &\overset{\text{i.i.d.}}{\sim} \pi \\ f \mid (A_1, \dots, A_d) &\sim W^A \end{aligned} \tag{5.3}$$

with π the horseshoe density on $\mathbb{R}_+^*$ and $W^A = \left\{ W_{(A_1 t_1, \dots, A_d t_d)} : \ t = (t_1, \dots, t_d) \in [-1,1]^d \right\}$ for W a squared exponential process.

We set $\pi = \pi_\tau$ as the *horseshoe* density (see, e.g., [26, 159]), i.e. the density of a random variable X_τ distributed as

$$\begin{aligned} \lambda &\sim C^+(0,1) \\ X_\tau \mid \lambda \ &\sim \mathcal{N}^+(0, \tau^2 \lambda^2), \end{aligned}$$

with $C^+(0,1)$ a standard half-Cauchy distribution and $\mathcal{N}^+(\mu, \sigma^2)$ is the half-normal distribution of $|X|$, $X \sim \mathcal{N}(\mu, \sigma^2)$.

The multi-layer case.

In order to perform some inference in this model, we introduce a Deep Gaussian Process-type prior, mixing ideas from [59] and the above introduction of the horseshoe distribution. We first place a prior Π on some parameters of the compositional structure such that $\Pi[q, d_1, \dots, d_q] = \Pi_q[q] \Pi[d_1, \dots, d_q | q]$. Given these parameters, we define a random regression function $f = g_q \circ \cdots \circ g_0$ where, for $i = 0, \dots, q$, $j = 1, \dots, d_{i+1}$, the applications g_{ij} are to $1 \wedge (h_{ij} \vee -1) =: \Psi(h_{ij})$ with h_{ij} independently distributed as in (5.3). Constraining the sample paths between -1 and 1 ensures that the composition is well-defined.

The *Deep Horseshoe Gaussian Process* $\mathrm{Deep\text{–}HGP}$ is defined as the hierarchical prior

$$\begin{aligned} q &\sim \Pi_q \\ d_1, \dots, d_q \mid q &\sim \Pi[\cdot | q] \\ A_{ij} \mid q, d_1, \dots, d_q &\overset{\text{i.i.d.}}{\sim} \pi_\tau^{\otimes d_i} \\ g_{ij} \mid q, d_1, \dots, d_q &\overset{\text{i.i.d.}}{\sim} W^{A_{ij}} \\ f \mid q, d_1, \dots, d_q, g_{ij} &= \Psi(g_q) \circ \cdots \circ \Psi(g_0). \end{aligned}$$

for $\tau > 0$.

5.3 Main results: deep simultaneous adaptation to structure and smoothness

5.3.1 Single layer setting: shallow horseshoe GP

The first theorem below is a special case of Theorem 25 and we present it to facilitate the reading of this chapter. We keep it to better emphasize the ability of HGP the assumption of smaller dimension than the ambient sampling space and to motivate future works on potential applications to high-dimensional settings (with d potentially greater than n).

Theorem 24. *Suppose f_0 is as in the first setting described in Section 5.2.1, for $d^* \leq d$ fixed. Let Π be the hierarchical distribution (5.3) with $\pi = \pi_\tau$. Then, there exists $\kappa > 0$, depending on d, such that for any $\beta > 0$ and $0 < \rho < 1$, $\Pi^\rho[\cdot \,|\, X, Y]$ contracts to f_0 at the rate $r_n = n^{-\frac{\beta}{2\beta+d^*}} \log^\kappa n$.*

5.3.2 Multilayer setting: deep horseshoe GP

The following results shows that the multilayer version of our prior adapts to the unknown compositional structure of the regression function. The fractional posterior attains the minimax rate of contraction, up to some polylog factor.

Theorem 25. *Suppose $f_0 \in \mathcal{F}(\lambda, \beta, K)$, with unknown parameters λ, β, K. Let Π be the Deep–HGP prior for arbitrary $\tau > 0$. Then, there exists a constant κ depending on the model parameters such that, for any $0 < \rho < 1$, $\Pi^\rho[\cdot \,|\, X, Y]$ contracts to f_0 at the rate $r_n = \tau_n^* \log^\kappa n$.*

5.4 Discussion

The use of fractional posteriors. We obtain our main results for fractional posterior distributions, where the parameter ρ can be taken to be any constant in $(0, 1)$. Two natural questions are: a) what is the interest in doing this; and b) what do we lose with respect to the usual posterior $\rho = 1$? The answer to a) lies in the need, when using the general theory of [64], to build sieves, capturing most prior mass, whose entropy or 'complexity' is well controlled. However, sieves are sometimes difficult to construct, especially if the probability of the complement of the sieve has to have a form of exponentially fast decrease. This difficulty leads [59] to condition sample paths of Gaussian processes to verify certain smoothness constraints. Although we do not have an answer to this specific question for $\rho = 1$, we note that this condition is not present to guarantee the convergence of the ρ–posterior, so we do not need to condition on boundedness of derivatives in our prior construction. This is an advantage also computationally, as adding more conditioning constraints may typically slow down MCMC samplers.

On the other hand, we argue that, at least for the set of applications of Bayesian (possibly tempered) posteriors considered here, one does not loose much, except perhaps in the constants in the convergence rate, but the latter is typically stated up to a large enough constant, so is not a concern at least for the theory. First, regarding sampling algorithms in practice, most sampling methods such as MCMC are of similar difficulty with the fractional or the original likelihood, so this is not a major concern computationally. One looses, though, the interpretation of the posterior as a conditional distribution, as well as the decision–theoretical back-up that comes with it, but also efficiency for $\sqrt{n}$–estimable parameters, that comes with the Bernstein–von Mises theorem, that will not hold as such for ρ–posteriors (this can be remedied under some conditions though, as investigated in [104]). However, again, this is not a main concern here, as we are mainly interested in nonparametric convergence rates up to constants.

Simulations. Although the computational aspect is not the focus of the present work, one main aim here is to provide (near)–optimal theoretical guarantees for a prior as simple as possible. We note that the prior distribution considered comes already fairly close to what is used in practice: indeed, note that given the lengthscale parameters, the deep GP prior corresponds

to the one considered in the original paper [44], where the kernel is termed ARD (Automatic Relevance Determination) and the lengthscales are called weights. In [44], the weights are then calibrated using a variational approach. Adapting this computational approach to the $\mathrm{Deep\text{–}HGP}$ is an interesting avenue for future work.

Work in progress and open questions. A natural open question of mathematical nature is whether the results obtained above carry over to original posteriors $\rho = 1$. As we have seen, this has no immediate consequence on simulations, but a positive answer would be particularly satisfactory. In addition, there is hope that HGP can provide near optimal results in a high-dimensional setting, with d much larger than n, which is an idea currently under investigation. Also, regarding the multilayer setting and $\mathrm{Deep\text{–}HGP}$, an extension of the theory presented here to the case where dimensions and at least one of the layer dimensions are allowed to grow with n is currently work in progress.

5.5 Proof of the main results

Reducing the problem to a prior mass condition.

Let us recall the definition of the α–Rényi divergence

$$R_\alpha(f,g) \coloneqq R_\alpha\left(P_f, P_g\right) = -\log \int p_f^\alpha p_g^{1-\alpha} d\mu, \quad \alpha > 0.$$

By Jensen's inequality, it follows that

$$R_\alpha(f,g) = -\log \int e^{-\frac{\alpha-\alpha^2}{2\sigma_0^2}(f-g)^2} d\mu \leq \frac{\alpha-\alpha^2}{2\sigma_0^2} \|f-g\|_{L_2(\mu)}^2 .$$

According to Theorem 3.1. in [93], the discussion below it and Example 8.44 in [65], for $KL(f,g) \coloneqq KL(P_f \mid\mid P_g)$ the Kullback-Leibler divergence between P_f and P_g, it is sufficient to have

$$\Pi\left[f:\ KL(f,f_0) < \varepsilon_n^2\right] \geq e^{-Cn\varepsilon_n^2},$$

for some $C > 0$ to ensure that ε_n is an R_α-posterior contraction rate for $\Pi^{(\rho)}$ in the sense that

$$\mathbb{E}_{f_0}\Pi^{(\rho)}\left[f:\ R_\alpha(f,f_0) \geq M_n\varepsilon_n \middle| X, Y\right] \to 0$$

for any $M_n \to \infty$. As explained above, this also translates into $L_2(\mu)$-contraction rates.

Since the supremum norm is a stronger metric than the Kullback-Leibler divergence and the standard theory for GPs considers more naturally this metric, we translate the above small ball condition. Standard arguments (see Section 8.1 of [59] for instance) indeed show that

$$\left\{f:\ \|f-f_0\|_\infty < \sqrt{2}\varepsilon\right\} \subset \left\{f:\ KL(f,f_0) < \varepsilon^2\right\}$$

so that in the following, we aim at finding lower bounds on

$$\Pi\left[f:\ \|f-f_0\|_\infty < \varepsilon_n\right].$$

5.5.1 Lower bound on the small ball probability.

Denoting by φ_g^A the concentration function corresponding to the process W^A from Section 5.2.3 and its RKHS $\mathbb{H}^A$, we have

$$P_{\varepsilon_n} := \Pi\left[f : \ \|f - f_0\|_\infty < \varepsilon_n\right] \geq \int \dots \int e^{-\varphi_{f_0}^A(\varepsilon_n)} \prod_{i=1}^d \pi(A_i) d\lambda(A_1) \dots d\lambda(A_d).$$

According to Lemma 40, for $\bar{A} := \max_i A_i > 0$ large enough, there exists absolute constants C, ε_0 such that whenever $\varepsilon < \varepsilon_0$,

$$-\log P\left[\left\|W^A\right\| < \varepsilon\right] \leq C^d d^{11d/2} \log^{1+d}(d) \left(\log \frac{\bar{A}}{\varepsilon}\right)^{1+d} \prod_i (1 \vee A_i). \tag{5.4}$$

Also, Lemma 39 ensures that, for f_0 β–Hölder, if $A_i = a > 0$ for $i \in \{i_1, \dots, i_{d^*}\}$, for some absolute constant $C > 0$, $D > 0$ depending on α, and any $\varepsilon \geq D d^* a^{-\beta}$ and $A_{i_1}, \dots, A_{i_{d^*}}$ large enough,

$$\inf_{\substack{h \in \mathbb{H}^{\mathbb{A}}: \\ \|h - f_0\|_\infty \leq \varepsilon}} \|h\|_{\mathbb{H}^{\mathbb{A}}}^2 \leq C^d \prod_{i=1}^d A_i \, \|f_0\|_2^2 \,. \tag{5.5}$$

Let's now write, for some $T > 0$,

$$a_i^* = \begin{cases} T, & \text{if } i \notin \{i_1, \dots, i_{d^*}\} \\ n^{\frac{1}{2\beta + d^*}}, & \text{otherwise.} \end{cases}$$

Combining (5.4) and (5.5), for some ε_n to be characterized, we lower bound P_{ε_n} by

$$\int_{a_1^*}^{2a_1^*} \dots \int_{a_d^*}^{2a_d^*} e^{-\varphi_{f_0}^A(\varepsilon_n)} \prod_{i=1}^d \pi(A_i) \, d\lambda(A_d) \dots d\lambda(A_1),$$

and on the domain of integration, (5.4) and (5.5) give, for C absolute constant,

$$\begin{aligned} \varphi_{f_0}^A(\varepsilon_n) &\leq (2C)^d \prod_i (1 \vee a_i^*) \left[d^{11d/2} \log^{1+d}(d) \left(\log \frac{\bar{A}}{\varepsilon_n}\right)^{1+d} + \|f_0\|_2^2 \right] \\ &\leq (2C)^d (1 \vee T)^{d - d^*} n^{\frac{d^*}{2\beta + d^*}} \left[d^{11d/2} \log^{1+d}(d) \left(\log \frac{n^{\frac{1}{2\beta + d^*}}}{\varepsilon_n}\right)^{1+d} + \|f_0\|_2^2 \right]. \end{aligned}$$

for any $\varepsilon_n \geq D d^* n^{-\frac{\beta}{2\beta + d^*}}$ and n large enough.

For d and d^* fixed, it is possible to take T constant and $\varepsilon_n = M n^{-\frac{\beta}{2\beta + d^*}} \log^c n$, $M, c > 0$ large enough, to verify $\varphi_{f_0}^A(\varepsilon_n) \leq n\varepsilon_n^2$.

Dimension-dependent upper bounds on the concentration function.

In order to prove posterior contraction rates for the GPs, we need to obtain upper bounds on the terms in the concentration function (5.2). The results below are essentially corolaries of Lemmas 4.2 and 4.3 from [9], with explicit constants depending on the ambient dimension d in the result. We indeed plan to apply the Deep–HGP prior in a high-dimensional setting

in future works, which requires to explicit the dependency on d of any quantities involved. As opposed to [9], we do not consider the anisotropic case, in which the function f_0 has different regularities depending on the direction. We focus on the variable selection aspect of the problem, assuming the same regularity for the directions on which f_0 depends.

Let's take W the squared-exponential stationary Gaussian process with covariance kernel $K(t) = \int e^{i\langle u,t\rangle} d\nu(u)$ and spectral measure ν that is the distribution of $d-$dimensional Gaussian random vector, with independent marginals of variance 2. We denote g the density of a standard random variable. The following lemma deals with the decentering part of the concentration function.

Lemma 39. *Suppose f_0 is as in the first setting of Section 5.2.1 and $A = (A_1, \dots, A_d)$ is such that $A_i = a > 0$ for $i \in \{i_1, \dots, i_{d^*}\}$. Let ψ be a higher order kernel such that*

- $\int \psi(t)dt = 1$;
- $\int t^k \psi(t)dt = 0$ *for* $k = 0, \dots, \lfloor \beta \rfloor$;
- $\int |t|^\beta |\psi(t)|dt < \infty$,

and its Fourier transform $\widehat{\psi}$ is compactly supported. Then,

$$\inf_{\substack{h \in \mathbb{H}^{\mathbb{A}}: \\ \|h - f_0\|_\infty \leq d^* \int |s|^\beta \psi(s) ds \; a^{-\beta}}} \|h\|^2_{\mathbb{H}^{\mathbb{A}}} \leq \prod_{i=1}^d A_i \left\| \frac{|\widehat{\psi}|^2}{g} \right\|^d \|f_0\|_2^2 .$$

Proof. We follow the lines of the proof of Lemma 4.2 from [9], giving the main elements.

Define $\tilde{\Psi} : \mathbb{R}^d \to \mathbb{C}$ by $\Psi(t) = \psi(t_1) \cdots \psi(t_d)$ and $\tilde{\Psi}_A(t) = \tilde{\Psi} A_1 t_1, \dots, A_d t_d) \prod_{i=1}^d A_i$. From the proof of Lemma 4.3 in [165], on which Lemma 4.2 [9] is based, the convolution $\Psi_A * f_0$ is an element of $\mathbb{H}^A$ with its square norm bounded by $\prod_{i=1}^d A_i \left\| \frac{|\widehat{\tilde{\Psi}}|^2}{\mathbf{g}} \right\| \|f_0\|^2_{L^2}$ with $\mathbf{g}$ the density of a multivariate standard normal gaussian vector. We note that

$$\left\| \frac{|\widehat{\tilde{\Psi}}|^2}{\mathbf{g}} \right\| \leq \left\| \frac{|\widehat{\psi}|^2}{g} \right\|^d < \infty$$

using the compactness of the support of $\widehat{\psi}$. It remains to show that this element approximate f_0 well enough, which follows from

$$\left| \tilde{\Psi}_A * f_0(t) - f_0(t) \right| \leq \left| \sum_{j=1}^{d^*} \int \psi(s_j) S_{i_j}(t_{i_j}, s_{i_j}/A_{i_j}) ds_{i_j} \right|, \quad t \in [0,1]^d,$$

with $\left| S_{i_j}(t_{i_j}, s_{i_j}/A_{i_j}) \right| \leq K \left| s_{i_j} \right|^\beta a^{-\beta}$ and K independent from the dimension d, as proved in the proof of Lemma 4.2 [9]. Then, we have proved that $\left\| \tilde{\Psi}_A * f_0 - f_0 \right\|_\infty \leq d^* \int |s|^\beta \psi(s) ds \, a^{-\beta}$ which concludes the proof. □

We now deal with the small ball probability in the concentration function and bound it. In the following, we note the ε-covering number $N(\varepsilon, S, D)$ of a semimetric space S equipped with a semimetric D as the minimal number of balls of radius ε needed to cover S.

Lemma 40. *For any $\bar{A} = \max_i A_i$ greater than an absolute constant, there exists an absolute constant $C, \varepsilon_0 > 0$ such that for $0 < \varepsilon < \varepsilon_0$,*

$$\varphi_0^A(\varepsilon) := -\log P\left[\left\|W^A\right\| < \varepsilon\right] \leq C^d d^{11d/2} \log^{1+d}(d) \left(\log \frac{1 \vee \bar{A}}{\varepsilon}\right)^{1+d} \prod_i (1 \vee A_i).$$

Proof. We follow the proof of Lemma 4.6 in [165]. We start with formula (3.25) in [94] which states that, for any $\varepsilon, \lambda > 0$,

$$\varphi_0^A(2\varepsilon) + \log \Phi\left(\lambda + \Phi^{-1}(e^{-\varphi_0^A(2\varepsilon)})\right) \leq \log\ N\left(\varepsilon, \mathbb{H}_1^A, \|\cdot\|_\infty\right),$$

with $\mathbb{H}_1^A$ the unit ball of $\mathbb{H}^A$ and Φ the standard normal distribution function. For the choice $\lambda = \sqrt{2\varphi_0^A(\varepsilon)}$ and with the inequality (see Lemma 4.10 in [165]) $\Phi\left(\sqrt{2x} + \Phi^{-1}(e^{-x})\right) \geq 1/2$, for any $x > 0$, we get

$$\varphi_0^A(2\varepsilon) + \log(1/2) \leq \log\ N\left(\varepsilon/\sqrt{2\varphi_0^A(\varepsilon)}, \mathbb{H}_1^A, \|\cdot\|_\infty\right). \tag{5.6}$$

Before going further, it is necessary to prove a crude bound of the form $\varphi_0^A(\varepsilon) \lesssim (\max_i A_i/\varepsilon)^\tau$, for some $\tau > 0$. Let u_A be the mapping associated to W^A considered in [106] and, as in this article, set

$$e_n(u_A) := \inf\left\{\eta > 0 :\ \log\ N\left(\eta, \mathbb{H}_1^A, \|\cdot\|_\infty\right) \leq (n-1)\log 2\right\}.$$

Lemma 41 implies that, for $\varepsilon > 0$,

$$\log\ N\left(\varepsilon, \mathbb{H}_1^A, \|\cdot\|_\infty\right) \leq C_d \log\left(\frac{1}{\varepsilon}\right)^{1+d} (1 \vee \bar{A})^d,$$

with C an absolute constant and

$$C_d = C^d d^{9d/2} \log(d)^{1+d}.$$

By definition, $e_n(u_A)$ is smaller than the solution η^* of $C_d \log\left(\frac{1}{\eta^*}\right)^{1+d} (1 \vee \bar{A})^d = (n-1)\log 2$, that is

$$\eta^* = e^{-C^{\frac{1}{1+d}} n^{\frac{1}{1+d}} C_d^{-\frac{1}{1+d}} \bar{A}^{-\frac{d}{1+d}}}.$$

We rewrite the first equation of [156] as $\sup_{k \leq n} k^\alpha e_n(u_A^*) \leq 32 \sup_{k \leq n} k^\alpha e_n(u_A)$ for any $n \geq 1, \alpha > 0$. Also, for any $k \geq 1$, since $x \to x e^{-x^{\frac{1}{1+d}}}$ is upper bounded by $(1+d)^{1+d} e^{-(1+d)}$ on $\mathbb{R}_+^*$,

$$\begin{aligned} k e_k(u_A) &\leq k e^{-C^{\frac{1}{1+d}} k^{\frac{1}{1+d}} C_d^{-\frac{1}{1+d}} (1 \vee \bar{A})^{-\frac{d}{1+d}}} \\ &= \frac{C_d (1 \vee \bar{A})^d}{C} \left(\frac{Ck}{C_d (1 \vee \bar{A})^d}\right) e^{-C^{\frac{1}{1+d}} k^{\frac{1}{1+d}} C_d^{-\frac{1}{1+d}} (1 \vee \bar{A})^{-\frac{d}{1+d}}} \\ &\leq \frac{C_d (1 \vee \bar{A})^d}{C} \left(\frac{1+d}{e}\right)^{1+d}. \end{aligned}$$

This implies that, for $n \geq 1$,

$$n e_n(u_A^*) \leq \sup_{k \leq n} k e_n(u_A^*)$$

$$\leq 32 \sup_{k\leq n} k e_n(u_A)$$
$$\leq 32 \frac{C_d(1\vee \bar{A})^d}{C}\left(\frac{1+d}{e}\right)^{1+d}.$$

From Lemma 2.1 [106], itself cited from [133], and this last upper bound, gives the following upper bound on $l_n(u_A)$ (defined in Section 2 of [106]): for c_1, c_2 absolute constants,

$$l_n(u_A) \leq c_1 \sum_{k\geq c_2 n} e_k(u_A^*) k^{-1/2}(1+\log k) \leq C C_d \bar{A}^d n^{-1/2},$$

for C an absolute constant independent.

From the proof of Proposition 2.4 in [106], we find that, for $\varepsilon > 0$, $\sigma = \mathbb{E}[\left\|W^A\right\|_\infty^2]^{1/2}$ and

$$n_A(\varepsilon) := \max\{n : \ 4l_n(u_A) \geq \varepsilon\},$$

the following bound stands

$$P\left[\left\|W^A\right\| < \varepsilon\right] \geq \frac{3}{4}\left(\frac{\varepsilon}{6\sigma n_A(\varepsilon)}\right)^{n_A(\varepsilon)},$$

which implies

$$\varphi_0^A(\varepsilon) \leq n(\varepsilon)\log\left(\frac{8\sigma n(\varepsilon)}{\varepsilon}\right). \tag{5.7}$$

We note that $n_A(\varepsilon)$ as long as ε is smaller than an absolute constant: indeed, $l_n(u_A)$ decreases with n and $l_1(u_A) = \sigma > \mathbb{E}[W_0^{A2}]^{1/2} = 1$. The above bound on $l_n(u_A)$ ensures that $n_A(\varepsilon) \leq C\left(C_d(1\vee\bar{A})^d\right)^2 \varepsilon^{-2}$.

Since $\sigma^2 = \mathbb{E}[\left\|W^A\right\|_\infty]^2 + \mathbb{V}[\left\|W^A\right\|_\infty]$, we bound these two terms. Since $\sup_{t\in[0,1]^d} \mathbb{E}W_{At}^2 = 1$, Theorem 2.5.8 [69] gives the bound on the tail probability $P\left(\left|\left\|W^A\right\|_\infty - \mathbb{E}\left\|W^A\right\|_\infty\right| > u\right) \leq 2e^{-u^2/2}$ and then

$$\mathbb{V}[\left\|W^A\right\|_\infty] = \int_0^\infty 2xP\left(\left|\left\|W^A\right\|_\infty - \mathbb{E}\left\|W^A\right\|_\infty\right| > x\right)dx \leq 4\int_0^\infty xe^{-x^2/2}dx = 4.$$

We control the other term via Theorem 2.3.7 of [69]:

$$\mathbb{E}[\left\|W^A\right\|_\infty] \leq \mathbb{E}|X| + 4\sqrt{2}\int_0^{M/2}\sqrt{\log(2N(\varepsilon,[0,1]^d,D))}d\varepsilon$$

for $X\sim\mathcal{N}(0,1)$, $D(s,t) = 2\left(1-e^{-\sum_i A_i^2(t_i-s_i)^2}\right)$ and $M = \sup\left\{s,t\in[-1,1]^d : \ D(s,t)\right\} = 2\left(1-e^{-4\sum_i A_i^2}\right)$. We note that $\|s-t\|_2 \leq \sqrt{-\log(1-\varepsilon/2)}/\bar{A}$ implies $D(s,t)\leq\varepsilon$ for $\varepsilon>0$, so, for $B_0(r)$ the euclidean ball of radius $r>0$ around 0,

$$N\left(\varepsilon,[-1,1]^d,D\right) \leq N\left(\sqrt{-\log(1-\varepsilon/2)}/\bar{A}, B_0(\sqrt{d}), \|\cdot\|_2\right) \leq \left(\frac{3\sqrt{d}\bar{A}}{\sqrt{-\log(1-\varepsilon/2)}}\right)^d$$

by standard arguments (see Proposition C.2 in [65]). The above integral is then bounded by, with the next inequalities involving absolute constants only,

$$\int_0^{1-e^{-4d\bar{A}^2}}\sqrt{\log(2N(\varepsilon,[0,1]^d,D))}d\varepsilon \lesssim \sqrt{d\log(d\bar{A})}\int_0^{1-e^{-4d\bar{A}^2}}\sqrt{\log\left(\frac{1}{\sqrt{-\log(1-\varepsilon/2)}}\right)}d\varepsilon$$

$$\lesssim \sqrt{d \log(d(1 \vee \bar{A}))}.$$

We conclude that $\sigma^2 \lesssim d \log(d(1 \vee \bar{A}))$.

Going back to (5.7), we conclude this proof with the bound, for $\bar{A} \geq 1$,

$$\begin{aligned}\varphi_0^A(\varepsilon) &\leq \left(C_d(1 \vee \bar{A})^d\right)^2 \varepsilon^{-2} \log\left(\sqrt{d \log(d(1 \vee \bar{A}))}\left(C_d(1 \vee \bar{A})^d\right)^2 \varepsilon^{-3}\right) \\ &\lesssim C^d d^{10d}(1 \vee \bar{A})^{2d} \log(d)^{3+2d} \log(1 \vee \bar{A}) \varepsilon^{-3},\end{aligned}$$

which we plug into (5.6) with Lemma 41, and C an absolute constant:

$$\varphi_0^A(\varepsilon) \leq C^d d^{11d/2} \log^{1+d}(d) \left(\log \frac{(1 \vee \bar{A})}{\varepsilon}\right)^{1+d} \prod_i (1 \vee A_i),$$

for $\varepsilon > 0$. □

Lemma 41. *For $\varepsilon > 0$ smaller than an absolute constant, and $\mathbb{H}_1^A$ the unit ball of $\mathbb{H}^A$, there exists absolute constants $C_1, C_2 > 0$ such that*

$$\log N\left(\varepsilon, \mathbb{H}_1^A, \|\cdot\|_\infty\right) \leq C_1^d d^{d/2} \left[d^2 \vee \log((C_2 d)^d/\varepsilon)\right]^d \left(\log \frac{(C_2 d)^d d^{1/4}}{\varepsilon}\right) \prod_i (1 \vee A_i).$$

Proof. We first note that ν has exponential moments, $\int e^{\delta \|t\|_2} d\nu(t) < \infty$ for $\delta > 0$. More precisely,

$$\int e^{\|t\|_2/2} d\nu(t) = \left(2^d \pi^{d/2}\right)^{-1} \int e^{\|t\|_2/2 - \|t\|_2^2/4} d\lambda(t) = \frac{1}{2^{d-1}\Gamma(d/2)} \int_0^{+\infty} r^{d-1} e^{r/2 - r^2/4} dr,$$

by change of variables in the integration of a radial function. The above integral is equal to

$$e^{1/4} \int_0^{+\infty} r^{d-1} e^{-\frac{(r-1)^2}{4}} dr,$$

and, splitting the domain of the integral, this is bounded by the sum of $e^{1/4} \int_0^1 r^{d-1} e^{-\frac{(r-1)^2}{4}} dr \leq e^{1/4}$ and

$$\begin{aligned}e^{1/4} \int_1^{+\infty} r^{d-1} e^{-\frac{(r-1)^2}{4}} dr &\leq e^{1/4} \int_0^{+\infty} (r+1)^{d-1} e^{-\frac{r^2}{4}} dr \\ &\leq e^{1/4} \sum_{i=0}^{d-1} \binom{d-1}{i} \int_0^{+\infty} r^i e^{-\frac{r^2}{4}} dr.\end{aligned}$$

The integrals in the above sum are, up to a universal constant factor, equal to the ith absolute moment of a truncated gaussian. As this distribution has subgaussian tails, these moments are upper bounded by $\left(K\sqrt{i}\right)^i$, for K a universal constant. So, via Stirling's formula,

$$C^2 := \int e^{\|t\|_2/2} d\nu(t) \leq \frac{(K\sqrt{d})^d}{\Gamma(d/2)} \leq \tilde{K}^d \sqrt{d},$$

where $\tilde{K}$ is a universal constant.

Let $k \in \mathbb{N}^*$. Adapting the proof of Lemma 4.4 in [165], a covering of $\mathbb{H}_1^A$ is build in the proof of Lemma 4.5 in [9], with less than, for c_1 an absolute constant,

$$\prod_{i=1}^{d} (A_i \vee 1) \, (c_1 d)^{d/2} \, k^d \log (2C/\varepsilon)$$

elements and $\|\cdot\|_\infty$-radii

$$C \sum_{l=k+1}^{\infty} \frac{l^{d-1}}{2^l} + \varepsilon \sum_{l=1}^{k} \frac{l^{d-1}}{2^l}.$$

We note that $\frac{l^{d-1}}{2^l} \leq (2/3)^l$ if $l \geq (d/\log(4/3))^2$, so that $\sum_{l=k+1}^{\infty} \frac{l^{d-1}}{2^l} \leq 2(2/3)^k \leq 2\varepsilon$, the last inequality being true for $k \geq \log(3/2)^{-1} \log(\varepsilon^{-1}) \vee (d/\log(4/3))^2$. Also, $\sum_{l=1}^{k} \frac{l^{d-1}}{2^l} \leq \sum_{l=1}^{\infty} (l+1)\ldots(l+d-1)2^{-l} \leq (d-1)!2^d$, so that, for k as above, we have a covering with radius, for c_2 an absolute constant independent of d,

$$\varepsilon \left(2C + (d-1)!2^d\right) \leq \varepsilon \left(2\tilde{K}^{d/2} d^{1/4} + (d-1)!2^d\right) \leq \varepsilon (c_2 d)^d.$$

We conclude that, for ε small enough, and c_3 an absolute constant,

$$\log \; N\left(\varepsilon, \mathbb{H}_1^A, \|\cdot\|_\infty\right) \leq c_3^d d^{d/2} \left[d^2 \vee \log(c_2 d)^d/\varepsilon\right]^d \log\left(\frac{2(c_2 d)^d \tilde{K}^d d^{1/4}}{\varepsilon}\right) \prod_i A_i,$$

which proves the assertion in the lemma. □

5.5.2 Proof of Theorem 24.

Given the developments at the beginning of Section 5.5 the discussion at the beginning of Section 5.5.1, it remains to bound the following volume for some $T > 0$

$$\left[\int_T^{2T} \pi_\tau(t)dt\right]^{d-d^*} \left[\int_{n^{\frac{1}{2\beta+d^*}}}^{2n^{\frac{1}{2\beta+d^*}}} \pi_\tau(t)dt\right]^{d^*}.$$

A lower bound is obtained with (5.9), which is for n large enough

$$e^{-Cd^* \log(n/\tau)}$$

for C an absolute constant. This implies that,

$$\Pi\left[f : \; \|f - f_0\|_\infty < r_n\right] \geq e^{-C\left(d^* n^{\frac{d^*}{2\beta+d^*}} + d^* \log\left(n^{\frac{d^{1+c_1} \log^{c_2} n}{d^*}}\right)\right)} \geq e^{-\tilde{C} n r_n^2}$$

which concludes the proof.

5.5.3 Proof of Theorem 25.

We follow the same arguments as in the above proof. When it comes to the prior mass, for any $f_0 \in \mathcal{F}(\lambda, \beta, K)$, we now have

$$\begin{aligned}\Pi\left[f : \; \|f - f_0\|_\infty < \varepsilon_n\right] \geq & \Pi_q[\{q\}]\Pi[\{d_1, \ldots, d_q\} |\; q] \\ & \Pi\left[\|\Psi(g_q) \circ \cdots \circ \Psi(g_0) - h_q \circ \cdots \circ h_0\|_\infty < \varepsilon_n \mid q, d_1, \ldots, d_q\right].\end{aligned}$$

Given the arguments from Section 5.5.1 and the fact that according to equation (5.9)

$$\left[\int_T^{2T} \pi(t)dt\right]^{d-d^*} \left[\int_{n^{\frac{1}{2\beta+d^*}}}^{2n^{\frac{1}{2\beta+d^*}}} \pi(t)dt\right]^{d^*} \geq e^{-Cd^* \log(n/\tau)},$$

we know that for $\varepsilon_n(\beta_i) = Mn^{-\frac{\beta_i}{2\beta_i+t_i}} \log^\kappa n$, $M, c > 0$ large enough,

$$\Pi\left[g_{ij} : \|\Psi(g_{ij}) - h_{ij}\|_\infty < Mn^{-\frac{\beta_i}{2\beta_i+t_i}} \log^c n \mid q, d_1, \ldots, d_q\right] \geq e^{-Cn\left(n^{-\frac{2\beta_i}{2\beta_i+t_i}}\right)\log^{2c} n},$$

where we used that $\|\Psi(g_{ij}) - h_{ij}\|_\infty \leq \|g_{ij} - h_{ij}\|_\infty$ as $\|h_{ij}\|_\infty \leq 1$. An application of Lemma 42 ensures that, for Q large enough,

$$\begin{aligned}
&\prod_{i=0}^{q}\prod_{j=1}^{d_{i+1}} \Pi\left[g_{ij} : \|g_{ij} - h_{ij}\|_\infty < Mn^{-\frac{\beta_i}{2\beta_i+t_i}} \log^c n \mid q, d_1, \ldots, d_q\right] \\
&\quad \leq \Pi\left[\|g_q \circ \cdots \circ g_0 - h_q \circ \cdots \circ h_0\|_\infty < Q\sum_{i=0}^{q}\left(n^{-\frac{\beta_i}{2\beta_i+t_i}} \log^c n\right)^{\alpha_i} \mid q, d_1, \ldots, d_q\right].
\end{aligned}$$

As a consequence, for C, κ large enough,

$$\Pi\left[f : \|f - f_0\|_\infty < C\tau_n^* \log^\kappa n\right] \geq e^{-Cn\tau_n^{*2}\log^{2\kappa} n},$$

and the $\rho-$posterior achieves minimax posterior contraction rates (up to some logarithmic factor).

Lemma 42. *Let $h_{ij} : [-1,1]^{d_i} \to [-1,1]$ be a function that depends on a subset S_{ij} of t_i coordinates and such that the restriction $h_{ij}\big|_{S_{ij}}$ satisfies $\left\|h_{ij}\big|_{S_{ij}}\right\|_{\beta_i,\infty} \leq K$ for some $\beta_i > 0$, $K \geq 1$. Then, the maps $h_i = (h_{ij})_{j=1,\ldots,d_{i+1}}^T$ satisfy for any $\tilde{h}_i = \left(\tilde{h}_{ij}\right)_{j=1,\ldots,d_{i+1}}^T$, with $\tilde{h}_{ij} : [-1,1]^{d_i} \to [-1,1]$,*

$$\left\|h_q \circ \ldots h_0 - \tilde{h}_q \circ \ldots \tilde{h}_0\right\|_{L^\infty[-1,1]^d} \leq K^q \sum_{i=0}^{q} \left\|\left\|h_i - \tilde{h}_i\right\|_\infty\right\|_\infty^{\alpha_i}$$

with $\alpha_i = \prod_{l=i+1}^q \beta_l \wedge 1$ (and $\alpha_q = 1$ by convention).

Proof. We follow the proof of Lemma 11 in [59] and prove the assertion by induction. For $q = 0$, this is trivially true. For $q = k+1 > 0$, assume that the statement is true for the positive integer k. We write $H_k = h_k \circ \ldots h_0$ and $\tilde{H}_k = \tilde{h}_k \circ \ldots \tilde{h}_0$ and use the triangle inequality so that

$$\begin{aligned}
&\left|h_{k+1} \circ H_k(x) - \tilde{h}_{k+1} \circ \tilde{H}_k(x)\right|_\infty \\
&\leq \left|h_{k+1} \circ H_k(x) - h_{k+1} \circ \tilde{H}_k(x)\right|_\infty + \left|h_{k+1} \circ \tilde{H}_k(x) - \tilde{h}_{k+1} \circ \tilde{H}_k(x)\right|_\infty \\
&\leq K\left|H_k(x) - \tilde{H}_k(x)\right|_\infty^{\beta_{k+1}\wedge 1} + \left|h_{k+1} \circ \tilde{H}_k(x) - \tilde{h}_{k+1} \circ \tilde{H}_k(x)\right|_\infty
\end{aligned}$$

$$
\begin{aligned}
&\leq K\left(K^k\sum_{i=0}^{k}\left\|\left|h_i-\tilde{h}_i\right|_\infty\right\|_\infty^{\prod_{l=i+1}^{k}\beta_l\wedge 1}\right)^{\beta_{k+1}\wedge 1}+\left\|\left|h_{k+1}-\tilde{h}_{k+1}\right|_\infty\right\|_\infty\\
&\leq K^{k+1}\sum_{i=0}^{k}\left\|\left|h_i-\tilde{h}_i\right|_\infty\right\|_\infty^{\prod_{l=i+1}^{k+1}\beta_l\wedge 1}+K^{k+1}\left\|\left|h_{k+1}-\tilde{h}_{k+1}\right|_\infty\right\|_\infty\\
&=K^{k+1}\sum_{i=0}^{k+1}\left\|\left|h_i-\tilde{h}_i\right|_\infty\right\|_\infty^{\alpha_i}
\end{aligned}
$$

where we use that $\left\|h_{(k+1)j}\right\|_{\beta_{k+1}\wedge 1,\infty}\leq K$ according to Lemma 10 of [147] and $(y+z)^\alpha\leq y^\alpha+x^\alpha$ for $y,z\geq 0$, $\alpha\in[0;1]$. □

5.6 The horseshoe density

Recall that π_τ denotes the horseshoe density defined in Section 5.2.3. Then, for any $t>0$,

$$
\begin{aligned}
\pi_\tau(t)&=\frac{2}{\pi}\frac{1}{\sqrt{2\pi\tau}}\int_{\mathbb{R}^+}\frac{1}{\lambda(1+\lambda^2)}e^{-\frac{t^2}{2\lambda^2\tau^2}}d\lambda\\
&\underset{v=\lambda^{-2}}{=}\frac{1}{\sqrt{2\pi^3\tau}}\int_{\mathbb{R}^+}\frac{1}{v+1}e^{-\frac{t^2}{2\tau^2}v}dv\\
&=\frac{e^{t^2/(2\tau^2)}}{\sqrt{2\pi^3\tau}}\underbrace{\int_1^{+\infty}\frac{1}{v}e^{-\frac{t^2}{2\tau^2}v}dv}_{=E_1\left(\frac{t^2}{2\tau^2}\right)}.
\end{aligned}
$$

It is known that (see Chapter 5 in [1]), for $x>0$,

$$
\frac{1}{2}e^{-x}\log\left(1+\frac{2}{x}\right)<E_1(x)<e^{-x}\log\left(1+\frac{1}{x}\right),
$$

so that we have the bound, for $t>0$,

$$
\frac{1}{(2\pi)^{3/2}\tau}\log\left(1+\frac{4\tau^2}{t^2}\right)<\pi_\tau(t)<\frac{1}{\sqrt{2\pi^3\tau}}\log\left(1+\frac{\tau^2}{t^2}\right). \tag{5.8}
$$

We refer to the prior Π_τ from (5.3) with $\pi=\pi_\tau$ as the Horseshoe Gaussian process in the following. From (5.8), we can lower bound, for n large enough and some absolute constant $C>0$,

$$
\left[\int_{n^{\frac{1}{2\beta+d^*}}}^{2n^{\frac{1}{2\beta+d^*}}}\pi_\tau(t)dt\right]^{d^*}\geq e^{-Cd^*\log(n/\tau)}. \tag{5.9}
$$

CHAPTER 6

Conclusion and perspectives

Concerning the questions laid out in 1.4, this thesis brought numerous parts of the solutions. The theoretical study of Bayesian tree-based methods carried out in Chapters 2 and 3 underlined their potential and flexibility. Despite their seemingly rough piecewise constant nature, ensembles of trees appear to provide optimal adaptive estimations in Hellinger distance for any regularities in density estimation. As shown in Chapter 2, forests can potentially outperform single-tree estimators as previous theoretical results limited themselves to optimality on classes with at most Lipschitz regularity. We drew a new connection with spline priors to demonstrate that these aggregation procedures can enjoy 'hidden' regularity structures. In Chapter 3, we showed that the OPT prior, based on a single tree structure, is enough to obtain optimal adaptive rates in the stronger supremum norm distance. In addition, we also built adaptive confidence sets of optimal radius in this norm, under additional self-similarity conditions on the signal. This answer to the question of uncertainty quantification for OPT posteriors advocates for using tree-based methods as full inferential machines. Then, as opposed to the supremum norm, we proved in Chapter 4 that the construction of confidence sets with optimal radius in Wasserstein distance, adapting to any regularities, exists in dimensions smaller than 4, without the need for additional qualitative assumptions (such as self-similarity). In higher dimensions, it is necessary to consider only values of regularity that are close enough but lying in a large window. Chapter 5 focused on Gaussian processes and their deep version and proved that optimal contraction rates could be achieved under structural assumptions of the signal.

The perspectives for the projects presented in this thesis include:

- obtaining optimality results on forest priors that are closer to the ones used in practice (e.g., BART);
- assuming the signal is more regular than Lipshitz, obtaining supremum contraction rates for forest priors, possibly via the definition of forests of OPTs, and developing results on uncertainty quantification for these priors;
- developing a construction of adaptive confidence sets in Wasserstein distance via Bayesian credible sets, as an alternative to the sample-splitting and risk estimation method;
- identifying the information-theoretic limitations of adaptive confidence sets in Wasserstein distances W_p, for $p > 2$;

- extending the results on the (Deep) Horseshoe Gaussian process to high-dimensional settings.

Bibliography

[1] Abramowitz, M. and Stegun, I. A., editors (1992). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover Publications, Inc., New York. Reprint of the 1972 edition.

[2] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR.

[3] Arlot, S. and Genuer, R. (2014). Analysis of purely random forests bias. *arXiv e-prints*, page arXiv:1407.3939.

[4] Armstrong, T. B. (2021). Adaptation bounds for confidence bands under self-similarity. *Bernoulli*, 27(2):1348–1370.

[5] Aubin, J.-P. (2000). *Applied functional analysis*. Pure and Applied Mathematics (New York). Wiley-Interscience, New York, second edition. With exercises by Bernard Cornet and Jean-Michel Lasry, Translated from the French by Carole Labrousse.

[6] Bachoc, F. and Lagnoux, A. (2021). Posterior contraction rates for constrained deep Gaussian processes in density estimation and classication. Arxiv preprint 2112.07280.

[7] Batir, N. (2008). Inequalities for the gamma function. *Arch. Math. (Basel)*, 91(6):554–563.

[8] Belitser, E. and Ghosal, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. volume 31, pages 536–559. Dedicated to the memory of Herbert E. Robbins.

[9] Bhattacharya, A., Pati, D., and Dunson, D. (2014). Anisotropic function estimation using multi-bandwidth gaussian processes. *The Annals of Statistics*, 42(1):352–381.

[10] Bhattacharya, A., Pati, D., and Yang, Y. (2019). Bayesian fractional posteriors. *Ann. Statist.*, 47(1):39–66.

[11] Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, 9:2015–2033.

[12] Biau, G., Sangnier, M., and Tanielian, U. (2021). Some theoretical insights into Wasserstein GANs. *J. Mach. Learn. Res.*, 22:Paper No. 119, 45.

[13] Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2):197–227.

[14] Biau, G., Scornet, E., and Welbl, J. (2019). Neural random forests. *Sankhya A*, 81(2):347–386.

[15] Birgé, L. (1984). Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.*, 3(2):259–282.

[16] Blanchard, G., Schäfer, C., and Rozenholc, Y. (2004). Oracle bounds and exact algorithm for dyadic classification trees. In *Learning theory*, volume 3120 of *Lecture Notes in Comput. Sci.*, pages 378–392. Springer, Berlin.

[17] Bleich, J. and Kapelner, A. (2014). Bayesian Additive Regression Trees With Parametric Models of Heteroskedasticity. *arXiv e-prints*, page arXiv:1402.5397.

[18] Boas, Jr., R. P. (1969). Inequalities for the derivatives of polynomials. *Math. Mag.*, 42:165–174.

[19] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[20] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA.

[21] Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *Ann. Statist.*, 30(4):927–961.

[22] Bull, A. (2012). Honest adaptive confidence bands and self-similar functions. *Electronic Journal of Statistics*, 6:1490–1516.

[23] Bull, A. D. and Nickl, R. (2013). Adaptive confidence sets in L^2. *Probab. Theory Related Fields*, 156(3-4):889–919.

[24] Cai, T. T. and Low, M. G. (2006). Adaptive confidence balls. *The Annals of Statistics*, 34(1):202–228.

[25] Carpentier, A. (2013). Honest and adaptive confidence sets in L_p. *Electron. J. Stat.*, 7:2875–2923.

[26] Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In van Dyk, D. and Welling, M., editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 73–80, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA. PMLR.

[27] Castillo, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.*, 2:1281–1299.

[28] Castillo, I. (2014). On Bayesian supremum norm contraction rates. *The Annals of Statistics*, 42:2058–2091.

[29] Castillo, I. (2017). Pólya tree posterior distributions on densities. *Ann. Inst. Henri Poincaré Probab. Stat.*, 53(4):2074–2102.

[30] Castillo, I., Kerkyacharian, G., and Picard, D. (2014). Thomas Bayes' walk on manifolds. *Probab. Theory Related Fields*, 158(3-4):665–710.

[31] Castillo, I. and Mismer, R. (2021). Spike and slab Pólya tree posterior densities: Adaptive inference. *Ann. Inst. Henri Poincaré Probab. Stat.*, 57(3):1521–1548.

[32] Castillo, I. and Nickl, R. (2013). Nonparametric Bernstein–von Mises Theorems in Gaussian white noise. *Ann. Statist.*, 41(4):1999–2028.

[33] Castillo, I. and Nickl, R. (2014). On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.*, 42(5):1941–1969.

[34] Castillo, I. and Ročková, V. (2021). Uncertainty quantification for Bayesian CART. *Ann. Statist.* to appear.

[35] Catoni, O. (2004). *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.

[36] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.*, 42(5):1787–1818.

[37] Chipman, H., George, E. I., Gramacy, R. B., and McCulloch, R. (2013). Bayesian treed response surface models. *Wiley Int. Rev. Data Min. and Knowl. Disc.*, 3(4):298–305.

[38] Chipman, H., George, E. I., and McCulloch, R. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4:266–298.

[39] Chipman, H., George, E. I., and McCulloch, R. E. (2000). Hierarchical priors for Bayesian CART shrinkage. *Statistics and Computing*, 10:17–24.

[40] Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART Model Search. *Journal of the American Statistical Association*, 93(443):935–948.

[41] Chipman, H. A., George, E. I., McCulloch, R. E., and Shively, T. S. (2016). High-dimensional nonparametric monotone function estimation using BART. *arXiv e-prints*, page arXiv:1612.01619.

[42] Christensen, J. and Ma, L. (2020). A Bayesian hierarchical model for related densities by using Pólya trees. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 82(1):127–153.

[43] Cohen, A., Daubechies, I., and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.*, 1(1):54–81.

[44] Damianou, A. and Lawrence, N. D. (2013). Deep Gaussian processes. In Carvalho, C. M. and Ravikumar, P., editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 207–215, Scottsdale, Arizona, USA. PMLR.

[45] Daubechies, I. (1992). *Ten lectures on wavelets*, volume 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

[46] De Boor, C. (1986). B(asic)-spline basics. Technical report, Wisconsin University - Madison mathematics research center.

[47] del Barrio, E., Giné, E., and Matrán, C. (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Ann. Probab.*, 27(2):1009–1071.

[48] Denil, M., Matheson, D., and Freitas, N. D. (2014). Narrowing the gap: Random forests in theory and in practice. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 665–673, Bejing, China. PMLR.

[49] Denison, D., Mallick, B., and Smith, A. (1998). A Bayesian CART algorithm. *Biometrika*, 85:363–377.

[50] Deshpande, S. K., Bai, R., Balocchi, C., Starling, J. E., and Weiss, J. (2020). VCBART: Bayesian trees for varying coefficients. *arXiv e-prints*, page arXiv:2003.06416.

[51] Ding, J., Bashashati, A., Roth, A., Oloumi, A., Tse, K., Zeng, T., Haffari, G., Hirst, M., Marra, M. A., Condon, A., Aparicio, S., and Shah, S. P. (2012). Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics (Oxford, England)*, 28(2):167–175.

[52] Donoho, D. L. (1997). CART and best-ortho-basis: a connection. *Ann. Statist.*, 25(5):1870–1911.

[53] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539.

[54] Dudley, R. M. (1968). The speed of mean Glivenko-Cantelli convergence. *Ann. Math. Statist.*, 40:40–50.

[55] Dudley, R. M. (2002). *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge. Revised reprint of the 1989 original.

[56] Duvenaud, D., Rippel, O., Adams, R. P., and Ghahramani, Z. (2014). Avoiding pathologies in very deep networks. In *Artificial Intelligence and Statistics*.

[57] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230.

[58] Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.*, 2:615–629.

[59] Finocchio, G. and Schmidt-Hieber, J. (2021). Posterior contraction for deep Gaussian process priors. *arXiv e-prints*, page arXiv:2105.07410.

[60] Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, 162(3-4):707–738.

[61] Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.*, 34:1386–1403.

[62] Freedman, D. A. (1965). On the asymptotic behavior of Bayes estimates in the discrete case. II. *Ann. Math. Statist.*, 36:454–456.

[63] Gey, S. and Nedelec, E. (2005). Model selection for CART regression trees. *IEEE Trans. Inform. Theory*, 51(2):658–670.

[64] Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.

[65] Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

[66] Giné, E. and Nickl, R. (2009). Uniform limit theorems for wavelet density estimators. *Ann. Probab.*, 37(4):1605–1646.

[67] Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.*, 38(2):1122–1170.

[68] Giné, E. and Nickl, R. (2011). Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist.*, 39(6):2883–2911.

[69] Giné, E. and Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York.

[70] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.

[71] Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). *Wavelets, approximation, and statistical applications*, volume 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

[72] He, J. and Hahn, P. R. (2020). Stochastic tree ensembles for regularized nonlinear regression. *arXiv e-prints*, page arXiv:2002.03375.

[73] He, J., Yalov, S., and Hahn, P. R. (2018). XBART: Accelerated Bayesian Additive Regression Trees. *arXiv e-prints*, page arXiv:1810.02215.

[74] Hill, J., Linero, A., and Murray, J. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7(1):251–278.

[75] Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.*, 20(1):217–240. Supplementary material available online.

[76] Hjort, N. L. and Walker, S. G. (2009). Quantile pyramids for Bayesian nonparametrics. *Ann. Statist.*, 37(1):105–131.

[77] Hoffmann, M. and Nickl, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.*, 39(5):2383–2409.

[78] Hoffmann, M., Rousseau, J., and Schmidt-Hieber, J. (2015). On adaptive posterior concentration rates. *The Annals of Statistics*, 43:2259–2295.

[79] Hoffmann-Jørgensen, J. (1974). Sums of independent Banach space valued random variables. *Studia Mathematica*, 52:159–186.

[80] Ibragimov, I. A. and Has′minskiĭ, R. Z. (1980). An estimate of the density of a distribution. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)*, 98:61–85, 161–162, 166. Studies in mathematical statistics, IV.

[81] Ioannou, Y., Robertson, D., Zikic, D., Kontschieder, P., Shotton, J., Brown, M., and Criminisi, A. (2015). Decision forests, convolutional networks and the models in-between. techreport MSR-TR-2015-58, Microsoft Research.

[82] Ishwaran, H. and Kogalur, U. B. (2010). Consistency of random survival forests. *Statist. Probab. Lett.*, 80(13-14):1056–1064.

[83] Jiang, H., Mu, J. C., Yang, K., Du, C., Lu, L., and Wong, W. H. (2016). Computational aspects of optional Pólya tree. *J. Comput. Graph. Statist.*, 25(1):301–320.

[84] Jiang, S. and Tokdar, S. T. (2021). Variable selection consistency of Gaussian process regression. *Ann. Statist.*, 49(5):2491–2505.

[85] Juditsky, A. and Lambert-Lacroix, S. (2003). Nonparametric confidence set estimation. *Math. Methods Statist.*, 12(4):410–428 (2004).

[86] Kantorovich, L. V. and Rubinshtein, S. (1958). On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, 13(7):52–59.

[87] Kantorovitch, L. (1942). On the translocation of masses. *C. R. (Doklady) Acad. Sci. URSS (N.S.)*, 37:199–201.

[88] Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41:495–502.

[89] Klusowski, J. M. (2018). Sharp Analysis of a Simple Model for Random Forests. *arXiv e-prints*, page arXiv:1805.02587.

[90] Klusowski, J. M. (2019). Analyzing CART. *arXiv e-prints*, page arXiv:1906.10086.

[91] Kontschieder, P., Fiterau, M., Criminisi, A., and Bulò, S. R. (2015). Deep Neural Decision Forests. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1467–1475.

[92] Kruijer, W., Rousseau, J., and van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.*, 4:1225–1257.

[93] Kruijer, W. and van der Vaart, A. (2013). Analyzing posteriors by the information inequality. In *From probability to statistics and back: high-dimensional models and processes*, volume 9 of *Inst. Math. Stat. (IMS) Collect.*, pages 227–240. Inst. Math. Statist., Beachwood, OH.

[94] Kuelbs, J. and Li, W. V. (1993). Metric entropy and the small ball problem for Gaussian measures. *J. Funct. Anal.*, 116(1):133–157.

[95] Lakshminarayanan, B., Roy, D. M., and Whye Teh, Y. (2014). Mondrian Forests: Efficient Online Random Forests. *arXiv e-prints*, page arXiv:1406.2673.

[96] Lamprinakou, S., McCoy, E., Barahona, M., Gandy, A., Flaxman, S., and Filippi, S. (2020). BART-based inference for Poisson processes. *arXiv e-prints*, page arXiv:2005.07927.

[97] Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.*, 20(3):1222–1235.

[98] Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York.

[99] Leahu, H. (2011). On the Bernstein-von Mises phenomenon in the Gaussian white noise model. *Electron. J. Stat.*, 5:373–404.

[100] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

[101] Lenk, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *J. Amer. Statist. Assoc.*, 83(402):509–516.

[102] Leonard, T. (1978). Density estimation, stochastic processes and prior information. *J. Roy. Statist. Soc. Ser. B*, 40(2):113–146. With discussion.

[103] Lepskii, O. V. (1991). On a Problem of Adaptive Estimation in Gaussian White Noise. *Theory of Probability & Its Applications*, 35(3):454–466.

[104] L'Huillier, A., Travis, L., Castillo, I., and Ray, K. (2022). Rates and shape of tempered posterior distributions. Manuscript in preparation.

[105] Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.*, 17(3):1001–1008.

[106] Li, W. V. and Linde, W. (1999). Approximation, metric entropy and small ball estimates for Gaussian measures. *Ann. Probab.*, 27(3):1556–1578.

[107] Linero, A. and Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Association*, 80:1087–1110.

[108] Linero, A. R. (2017). A review of tree-based Bayesian methods. *Communications for Statistical Applications and Methods*, 24(6):543–559.

[109] Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *J. Amer. Statist. Assoc.*, 113(522):626–636.

[110] Liu, L., Li, D., and Wong, W. H. (2017). Convergence rates of a partition based Bayesian multivariate density estimation method. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[111] Liu, Y., Shao, Z., and Yuan, G.-C. (2010). Prediction of polycomb target genes in mouse embryonic stem cells. *Genomics*, 96(1):17 – 26.

[112] Low, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.*, 25(6):2547–2554.

[113] Ma, L. (2017). Adaptive shrinkage in Pólya tree type models. *Bayesian Anal.*, 12(3):779–805.

[114] Ma, L. and Wong, W. H. (2011). Coupling optional Pólya trees and the two sample problem. *J. Amer. Statist. Assoc.*, 106(496):1553–1565.

[115] Matthews, A. G. d. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. (2018). Gaussian Process Behaviour in Wide Deep Neural Networks. *arXiv e-prints*, page arXiv:1804.11271.

[116] Mauldin, R. D., Sudderth, W. D., and Williams, S. C. (1992). Pólya trees and random distributions. *Ann. Statist.*, 20(3):1203–1221.

[117] McDonald, S. and Campbell, D. (2021). A review of uncertainty quantification for density estimation. *Stat. Surv.*, 15:1–71.

[118] Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.*, 17:Paper No. 26, 41.

[119] Meyer, Y. (1993). *Wavelets and Operators*, volume 1 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge.

[120] Monge, G. (1781). *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale.

[121] Mourtada, J., Gaïffas, S., and Scornet, E. (2020). Minimax optimal rates for Mondrian trees and forests. *Ann. Statist.*, 48(4):2253–2276.

[122] Naulet, Z. (2021). Adaptive Bayesian density estimation in sup-norm. arXiv preprint 1805.05816, to appear in *Bernoulli*.

[123] Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

[124] Ni, K., Bresson, X., Chan, T., and Esedoglu, S. (2009). Local histogram based segmentation using the wasserstein distance. *International Journal of Computer Vision*, 84(1):97–111.

[125] Nickl, R. and Ray, K. (2020). Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. *Ann. Statist.*, 48(3):1383–1408.

[126] Nickl, R. and Szabó, B. (2016). A sharp adaptive confidence ball for self-similar functions. *Stochastic Processes and their Applications*, 126(12):3913–3934.

[127] Nieto-Barajas, L. E. and Müller, P. (2012). Rubbery Polya tree. *Scand. J. Stat.*, 39(1):166–184.

[128] O'Reilly, E. and Tran, N. (2020). Stochastic geometry to generalize the Mondrian Process. *arXiv e-prints*, page arXiv:2002.00797.

[129] Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annu. Rev. Stat. Appl.*, 6:405–431.

[130] Pati, D., Bhattacharya, A., and Cheng, G. (2015). Optimal Bayesian estimation in random covariate design with a rescaled Gaussian process prior. *J. Mach. Learn. Res.*, 16:2837–2851.

[131] Peyré, G. and Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607.

[132] Picard, D. and Tribouley, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.*, 28(1):298–335.

[133] Pisier, G. (1989). *The volume of convex bodies and Banach space geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge.

[134] Pratola, M. T. (2016). Efficient Metropolis-Hastings proposal mechanisms for Bayesian regression tree models. *Bayesian Anal.*, 11(3):885–911.

[135] Pratola, M. T., Chipman, H. A., Gattiker, J. R., Higdon, D. M., McCulloch, R., and Rust, W. N. (2014). Parallel Bayesian additive regression trees. *J. Comput. Graph. Statist.*, 23(3):830–852.

[136] Pratola, M. T., Chipman, H. A., George, E. I., and McCulloch, R. E. (2020). Heteroscedastic BART via multiplicative regression trees. *J. Comput. Graph. Statist.*, 29(2):405–417.

[137] Randrianarisoa, T. (2021). Smoothing and adaptation of shifted Pólya Tree ensembles. arXiv preprint 2010.12299, to appear in *Bernoulli*.

[138] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.

[139] Ray, K. (2017). Adaptive Bernstein–von Mises theorems in Gaussian white noise. *The Annals of Statistics*, 45:2511–2536.

[140] Robins, J. and van der Vaart, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.*, 34(1):229–253.

[141] Rockova, V. (2019). On Semi-parametric Bernstein-von Mises Theorems for BART. *arXiv e-prints*, page arXiv:1905.03735.

[142] Rockova, V. and Rousseau, J. (2021). Ideal Bayesian Spatial Adaptation. *arXiv e-prints*, page arXiv:2105.12793.

[143] Rousseau, J. and Szabo, B. (2020). Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors. *Ann. Statist.*, 48(4):2155–2179.

[144] Ročková, V. and Saha, E. (2019). On Theory for BART. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research*, volume 89, pages 2839–2848. PMLR.

[145] Ročková, V. and van der Pas, S. (2020). Posterior concentration for Bayesian regression trees and forests. *Ann. Statist.*, 48(4):2108–2131.

[146] Roy, D. M. and Teh, Y. W. (2009). The Mondrian process. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 1377–1384. Curran Associates, Inc.

[147] Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist.*, 48(4):1875–1897.

[148] Scornet, E. (2016a). On the asymptotics of random forests. *J. Multivariate Anal.*, 146:72–83.

[149] Scornet, E. (2016b). Random forests and kernel methods. *IEEE Trans. Inform. Theory*, 62(3):1485–1500.

[150] Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *Ann. Statist.*, 43(4):1716–1741.

[151] Sethi, I. (1990). Entropy nets: from decision trees to neural networks. Technical report, Dept. of Computer Science, Wayne State Univ., Detroit, MI.

[152] Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714.

[153] Szabó, B., van der Vaart, A. W., and van Zanten, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.*, 43(4):1391–1428.

[154] Tan, Y. V. and Roy, J. (2019). Bayesian additive regression trees and the general BART model. *Stat. Med.*, 38(25):5048–5069.

[155] Teckentrup, A. L. (2020). Convergence of Gaussian process regression with estimated hyper-parameters and applications in Bayesian inverse problems. *SIAM/ASA J. Uncertain. Quantif.*, 8(4):1310–1337.

[156] Tomczak-Jaegermann, N. (1987). Dualité des nombres d'entropie pour des opérateurs à valeurs dans un espace de Hilbert. *C. R. Acad. Sci. Paris Sér. I Math.*, 305(7):299–301.

[157] 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.

[158] van der Pas, S. and Ročková, V. (2017). Bayesian dyadic trees and histograms for regression. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 2089–2099. Curran Associates, Inc.

[159] van der Pas, S. L., Kleijn, B. J. K., and van der Vaart, A. W. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8(2):2585 – 2618.

[160] van der Vaart, A. and van Zanten, H. (2007). Bayesian inference with rescaled Gaussian process priors. *Electron. J. Stat.*, 1:433–448.

[161] van der Vaart, A. and van Zanten, H. (2011). Information rates of nonparametric Gaussian process methods. *J. Mach. Learn. Res.*, 12:2095–2119.

[162] van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

[163] van der Vaart, A. W. and van Zanten, J. H. (2008a). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3):1435–1463.

[164] van der Vaart, A. W. and van Zanten, J. H. (2008b). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, volume 3 of *Inst. Math. Stat. (IMS) Collect.*, pages 200–222. Inst. Math. Statist., Beachwood, OH.

[165] van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675.

[166] Villani, C. (2009). *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin. Old and new.

[167] Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.*, 113(523):1228–1242.

[168] Wager, S. and Walther, G. (2015). Adaptive Concentration of Regression Trees, with Application to Random Forests. *arXiv e-prints*, page arXiv:1503.06388.

[169] Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B*, 40(3):364–372.

[170] Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648.

[171] Weed, J. and Berthet, Q. (2019). Estimation of smooth densities in Wasserstein distance. *arXiv:1902.01778 [math, stat]*.

[172] Welbl, J. (2014). Casting random forests as artificial neural networks (and profiting from it). In *Pattern recognition*, volume 8753 of *Lecture Notes in Comput. Sci.*, pages 765–771. Springer, Cham.

[173] Wong, W. H. and Ma, L. (2010). Optional Pólya tree and Bayesian inference. *Ann. Statist.*, 38(3):1433–1459.

[174] Yang, Y. and Dunson, D. B. (2016). Bayesian manifold regression. *The Annals of Statistics*, 44(2):876 – 905.

[175] Yang, Y., Morillo, I. G., and Hospedales, T. M. (2018). Deep neural decision trees. In *ICML Workshop on Human Interpretability in Machine Learning (WHI)*.

[176] Yang, Y. and Tokdar, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.*, 43(2):652–674.

[177] Yoo, W. and Ghosal, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *The Annals of Statistics*, 44(3):1069–1102.

[178] Zhang, J. L. and Härdle, W. K. (2010). The Bayesian additive classification tree applied to credit risk modelling. *Comput. Statist. Data Anal.*, 54(5):1197–1205.

[179] Zhang, T. (2006). From ϵ-entropy to KL-entropy: analysis of minimum information complexity density estimation. *Ann. Statist.*, 34(5):2180–2210.

Contributions to the theoretical analysis of statistical learning and uncertainty quantification methods

Thibault Randrianarisoa

Abstract

Modern data analysis provides scientists with statistical and machine learning algorithms with impressive performance. In front of their extensive use to tackle problems of constantly growing complexity, there is a real need to understand the conditions under which algorithms are successful or bound to fail. An additional objective is to gain insights into the design of new algorithmic methods able to tackle more innovative and challenging tasks. A natural framework for developing a mathematical theory of these methods is *nonparametric inference*. This area of Statistics is concerned with inferences of unknown quantities of interest under minimal assumptions, involving an infinite-dimensional statistical modeling of a parameter on the data-generating mechanism. In this thesis, we consider both problems of function estimation and uncertainty quantification.

The first class of algorithms we deal with are Bayesian tree-based methods. They are based on a 'divide-and-conquer' principle, partitioning a sample space to estimate the parameter locally. In regression, these methods include BCART and the renowned BART, the later being an ensemble of trees or a forest. In density estimation, the famous Pólya Tree prior exemplifies these methods and is the building block of a myriad of related constructions. We propose a new extension, DPA, that is a 'forest of PTs' and is shown to attain minimax contraction rates adaptively in Hellinger distance for arbitrary Hölder regularities. Adaptive rates in the stronger supremum norm are also obtained for the flexible Optional Pólya Tree (OPT) prior, a BCART-type prior, for regularities smaller than one.

Gaussian processes are another popular class of priors studied in Bayesian nonparametrics and Machine Learning. Motivated by the ever-growing size of datasets, we propose a new horseshoe Gaussian process with the aim to adapt to leverage a data structure of smaller dimension. First, we derive minimax optimal contraction rates for its tempered posterior. Secondly, deep Gaussian processes are Bayesian counterparts to the famous deep neural networks. We prove that, as a building block in such a deep framework, it also gives optimal adaptive rates under compositional structure assumptions on the parameter.

As for uncertainty quantification (UQ), Bayesian methods are often praised for the principled solution they offer with the definition of credible sets. We prove that OPT credible sets are confidence sets with good coverage and size (in supremum norm) under qualitative self-similarity conditions. Moreover, we conduct a theoretical study of UQ in Wasserstein distances W_p, uncovering a new phenomenon. In dimensions smaller than 4, it is possible to construct confidence sets whose W_p-radii, $p \geq 2$, adapt to any regularities (with no qualitative assumptions). This starkly contrasts the usual L_p theory, where concessions always have to be made.

Keywords: Bayesian nonparametrics, Tree-based methods, Uncertainty Quantification, Wasserstein distance, Gaussian process

LPSM
LABORATOIRE DE PROBABILITÉS STATISTIQUE & MODÉLISATION