UNIVERSITÉ DU QUÉBEC À MONTRÉAL

INFÉRENCE BAYÉSIENNE NONPARAMÉTRIQUE POUR LES DENSITÉS À SUPPORT MÉTRIQUE COMPACT ET PROBLÈMES APPARENTÉS

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN MATHÉMATIQUES

PAR
OLIVIER BINETTE

JUILLET 2019

REMERCIEMENTS

Je dédicace ce mémoire à ma famille, à mes amis, mentors et collaborateurs. J'espère que vous saurez vous y reconnaitre. C'est grâce à votre soutien sans réserve que j'ai aujourd'hui la liberté d'explorer le riche univers des probabilités et de ses applications à la compréhension de notre monde.

CONTENTS

LIST OF TABLES

LISTE DES FIGURES

RÉSUMÉ

Le thème principal de ce mémoire est l'estimation de densités définies sur des espaces métriques compacts en utilisant des méthodes bayésiennes nonparamétriques (Binette and Guillotte, 2018). Le cas où l'espace métrique est le cercle, d'intérêt en statistique circulaire et directionnelle, est développé avec une attention particulière. Nous proposons dans ce contexte une base de densités de probabilités des polynômes trigonométriques possédant des propriétés de préservation de la forme analogues aux densités polynomiales de Bernstein. Une étude de simulation montre que des estimateurs bayésiens nonparamétriques développés à l'aide de cette base peuvent offrir des gains par rapport à des méthodes comparables précédemment suggérées dans la littérature.

D'un point de vue théorique, nous étudions les propriétés de concentration, pour la distance de Hellinger, des distributions a posteriori issues de modèles engendrés par des opérateurs d'approximation linéaires positifs de rang fini. Ce type de modèles généralise les polynômes aléatoires de Bernstein à l'utilisation d'autres types de bases de densités de probabilités définies sur des espaces métriques compacts arbitraires. Ceux-ci se prêtent particulièrement bien à l'estimation sous contraintes de formes et les calculs a posteriori peuvent généralement être effectués à l'aide du Slice Sampler de Kalli et al. (2011). Nous obtenons la convergence de la distribution a posteriori sous des conditions de régularité particulièrement faibles ne nécessitant pas d'hypothèses de continuité. Des vitesses de convergences adaptatives sont de plus obtenues en termes de la croissance du rang des opérateurs et de leurs propriétés d'approximation.

Ces contributions sont liées à quelques bases mathématiques présentées dans le premier chapitre. Nous y introduisons différentes fonctions connues de divergences sur des ensembles de mesures de probabilités ainsi que leur relation au rapport de vraisemblance. De nouvelles inégalités de type *Pinsker inverse*, permettant d'obtenir des bornes optimales sur les f-divergences en termes de la variation totale et des extremums du rapport de vraisemblance (Binette, 2019), sont dérivées dans le Chapitre 2.

ABSTRACT

This work is concerned with density estimation on compact metric spaces using sieve priors (Binette and Guillotte, 2018). Particular attention is given to the case where the metric space is the circle as the problem is relevant to circular and directional statistics. In this context, we suggest a density basis of the trigonometric polynomials that is analogous, because of its interpretability and shape-preserving properties, to the Bernstein polynomial densities. A simulation study shows that the use of Bayes estimators constructed using this basis may provide gains over comparable circular density estimators previously suggested in the literature.

From a theoretical point of view, we study the convergence of posterior distribution, in the Hellinger sense, for models that arise as the images of positive linear approximation operators with finite ranks. These models generalize random Bernstein polynomials to the use of other density bases defined on arbitrary compact metric spaces. They are particularly well suited to shape constrained density estimation and posterior simulation may be carried out using the Slice Sampler of Kalli et al. (2011). Strong posterior consistency is obtained under notably weak regularity assumptions and adaptative convergence rates are expressed in terms of the growth of the operator ranks and of their approximation properties.

Some mathematical background is introduced in the first chapter. We introduce different known divergence functions over sets of probability measures as well as their relationship to the likelihood ratio. New *reverse Pinsker* inequalities, providing optimal upper bounds on f-divergences in terms of the total variation and likelihood ratio extremums (Binette, 2019), are derived in Chapter 2.

INTRODUCTION

Suppose that some unknown mechanism iteratively generates data points X_1, X_2, X_3, and so on. Our goal is to use finitely many of those observations, say $(X_i)_{i=1}^n$, to infer characteristics of the mechanism that may be of interest.

The way in which the points X_i are generated can be arbitrarily complex and may stochastically depend on external factors. The starting point of any meaningful statistical analysis would therefore be an assessment of the dependencies involved.

In the simplest case, which still abstractly encompasses a number of more general situations, we model the points X_i as random variables that are independent and identically distributed following some unknown probability distribution P_0.

Our *epistemic uncertainty* about what may be P_0 is quantified through what is called a *prior* probability distribution Π over the set of all reasonable possibilities for what P_0 may be. Given a set A of probability distributions, the prior probability $\Pi(A)$ of A specifies what we consider as the probability that "$P_0 \in A$" before any observation of the X_i has been made.

Once we have observed the data points $(X_i)_{i=1}^n$, we may adjust our prior quantification of uncertainty about P_0 through probabilistic conditioning, thus obtaining what is called the *posterior distribution* $A \mapsto \Pi\left(A \mid (X_i)_{i=1}^n\right)$.

This process of first quantifying uncertainty over an unknown state of the world through a prior probability measure and then making adjustments using the cal-

culus of probabilities in light of new observations is called *Bayesian inference.*

0.1 Subjects of this memoir

0.1.1 Metrics and divergences on probability measures

Our first chapter introduces some mathematical ideas relevant to the theoretical developments of the following chapters. We discuss different metrics and topologies on the space $\mathcal{M}$ of all probability measures on the space $\mathbb{M}$ on which the variables X_i take values, as this is related to the definition of prior distributions on $\mathcal{M}$ and to the study of properties of the posterior distributions.

0.1.2 Information inequalities

In chapter 2, we derive new best-possible inequalities allowing us to upper bound f-divergences in terms of the total variation distance and of likelihood ratio extremums. This work can be inscribed in the field of *Information Inequalities*: this is about relating together different distributional characteristics of the log likelihood ratio.

The motivation comes from Bayes' Theorem, which states that, in dominated models, the posterior distribution $\Pi(\cdot \mid (X_i)_{i=1}^n)$, for independent observations X_i with density f_0, may be written as

$$\Pi(A \mid (X_i)_{i=1}^n) \propto \int_A \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} \Pi(df). \tag{0.1.1}$$

The two elements involved in the right-hand side of this formula are the prior distribution Π and the likelihood ratio f/f_0. The logarithm of this likelihood ratio is commonly referred to as the "information" function $\iota(x) = \log \frac{f(x)}{f_0(x)}$ and we are interested in its distribution for x a random variable with density f_0.

The study of the behaviour of posterior distributions is therefore typically based on

properties of Π, on characteristics of the distribution of the information $\iota(x)$ over the range $f \in \mathbb{F}$ and on the resulting geometry on $\mathbb{F}$. Characteristics of ι include the total variation between f_0 and f, their Renyi divergence and their Kullback-Leibler divergence. Each may be expressed as an expected convex transform of $\exp \iota(x)$; they are what are called f-divergences.

Information inequalities relate together different characteristics of ι as well as the resulting geometries on $\mathbb{F}$. They are quite fundamental to the study of posterior distributions. One such inequality, of which we make repeated uses in Chapter 3, is

$$\int f_0 \log \frac{f_0}{f} \leqslant \left(\sup \frac{f_0}{f} \right) \int |f - f_0|.$$

This is an instance of a *reverse Pinsker inequality*: it upper bounds an f-divergence in terms of the total variation distance and the extremums of the likelihood ratio. While the above is immediate and already quite useful, it can be significantly improved. In Chapter 2, it is shown that any f-divergence D_ϕ, here characterized by a convex function $\phi : [0, \infty) \to \mathbb{R}$ with $\phi(1) = 0$ and

$$D_\phi(f_0, f) = \mathbb{E}\left[\phi\left(\frac{f(x)}{f_0(x)} \right) \right], \quad x \sim f_0,$$

we have

$$\sup_{(f_0, f) \in \mathcal{A}(m, M, \delta)} D_\phi(f_0, f) = \delta \left(\frac{\phi(m)}{1 - m} + \frac{\phi(M)}{M - 1} \right)$$

when considering the class $\mathcal{A}(m, M, \delta)$ of pairs (f_0, f) satisfying $\inf f/f_0 = m$, $\sup f/f_0 = M$ and $\int |f - f_0| = 2\delta$. This idea is developed in Binette (2019) as a response to suboptimal particular cases that appeared in the information inequalities literature.

0.1.3 Density estimation using sieve priors

Chapter 3 considers in some generality the case where the variables X_i take values in a compact metric space $(\mathbb{M}, d)$. For instance, the variables X_i may be

observations of angles distributed on the sphere or of directions distributed on a sphere.

We exploit sequences $T_n : L^1(\mathbb{M}) \to L^1(\mathbb{M})$, $n \in \mathbb{N}$, of positive linear operators with finite ranks mapping $\mathbb{F}$ to $\mathbb{F}$ in order to obtain the decomposition

$$\mathbb{F} = \overline{\bigcup_{n\in\mathbb{N}} T_n(\mathbb{F})},$$

where the overline denotes L^1 closure in $\mathbb{F}$. Given prior distributions Π_n on the submodels $T_n(\mathbb{F})$ and a distribution ρ on $\mathbb{N}$, we thus obtain a prior Π on $\mathbb{F}$ through

$$\Pi = \sum_{n\in\mathbb{N}} \rho(n)\Pi_n. \tag{0.1.2}$$

This is an instance of a *sieve prior* or *mixture prior* and a number of particular cases have appeared previously in the literature. Let me showcase a few examples and explain how some of our general results of Chapter 3 can easily be used to obtain asymptotic properties of the posterior distribution in terms of simple properties of the operators T_n.

Random Bernstein polynomials.

With $\mathbb{M} = [0, 1]$, take T_n the *Bernstein-Kantorovich* operator defined as

$$T_n f : x \mapsto (n+1) \sum_{i=0}^{n} \int_{\frac{i}{n+1}}^{\frac{i+1}{n+1}} f(u)\, du\, p_{i,n}(x),$$

where $p_{i,n}(x) = \binom{n}{i} x^i (1-x)^{n-i}$ is the ith Bernstein polynomial of degree n. It follows that

$$T_n(\mathbb{F}) = \left\{ (n+1) \sum_{i=0}^{n} c_{j,n} p_{j,n} : c_{j,n} \geqslant 0, \sum_j c_{j,n} = 1 \right\}$$

is the set of finite mixture of Bernstein polynomial densities of degree n. With Π_n a Dirichlet distribution on the coefficients $(c_{j,n})$ with parameters for instance $\alpha_{j,n} = 1/n$ and ρ a distribution on $\mathbb{N}$ with subexponential tail, we obtain a particular

case of the random Bernstein polynomials of Petrone (1999). Theorem 3.4.3 entail strong posterior consistency at all bounded densities provided also that $\rho(n) > 0$ for every $n \in \mathbb{N}$. Theorem 3.4.4, together with the well-known fact that $\|T_n f - f\|_\infty = \mathcal{O}\left(\omega_f(n^{-1/2})\right)$ where ω_f is the modulus of continuity of f, yields the posterior contraction rate $\varepsilon_n = (n/\log(n))^{-\beta/(2\beta+2)}$ whenever $\log \rho(n) \asymp -n \log n$ and the data generating density f_0 satisfies the the Hölder continuity condition $\omega_{f_0}(\delta) \leqslant C\delta^\beta$ for some $C > 0$ and $\beta > 0$. The strong posterior consistency result refines Theorem 2 of Petrone and Wasserman (2002) by removing continuity assumptions on f_0 while the posterior convergence rate we obtain is the same, up to log factors, as that obtained in Kruijer and van der Vaart (2008). However, the generality of our approach makes it readily applicable in other contexts as well.

Random histograms on metric spaces. Let $\mathbb{M}$ be an arbitrary compact metric space and let $\{R_{j,n}\}_{j=0}^{d_n}$ be a measurable partition of $\mathbb{M}$ of diameter less than n^{-1} with $d_n \in \mathbb{N}$ elements. We assume that $\max_j \mu(R_{j,n})^{-1} = \mathcal{O}(d_n)$ and that d_n is an increasing integer sequence satisfying $d_n \asymp n^d$ for some $d > 0$. Define

$$T_n f : x \mapsto \sum_{j=0}^{d_n} \int_{R_{j,n}} f(u) du \, \mu(R_{j,n})^{-1} \mathbb{1}_{R_{j,n}}(x)$$

so that

$$T_n(\mathbb{F}) = \left\{ \sum_{j=0}^{d_n} c_{j,n} \, \mu(R_{j,n})^{-1} \mathbb{1}_{R_{j,n}} : c_{j,n} \geqslant 0, \sum_j c_{j,n} = 1 \right\}$$

and $\|T_n f - f\|_\infty = \mathcal{O}(\omega_f(n^{-1}))$ for any continuous f. With Π_n a Dirichlet distribution on the coefficients $(c_{j,n})$ with parameters $\alpha_{j,n} = 1/d_n$ and ρ a distribution on $\mathbb{N}$ satisfying $\log \rho(n) \asymp -d_n \log(d_n)$, we obtain from Theorems 3.4.3 and 3.4.4 strong posterior consistency at any bounded density and the posterior convergence rate $\varepsilon_n = (n/\log(n))^{-\beta/(2\beta+d)}$ provided that f_0 satisfies a Hölder continuity condition with exponent $\beta > 0$.

The general form of positive linear operators. The Bernstein polynomials

and piecewise constant functions can be replaced by many other types of basis functions: splines with fixed knots, Gaussian kernels at predetermined locations, etc. The properties of resulting posterior distributions are then studied through the associated sequence of positive linear operators.

By the Riesz representation Theorem (Rudin, 1987), any positive linear operator and such that $T_n(1) = 1$ takes the form

$$T_n f : x \mapsto \mathbb{E}\left[f(Y_n(x))\right] \tag{0.1.3}$$

for some families $\{Y_n(x) : x \in \mathbb{M}\}$ of random variables. In the examples considered above, it is an easy exercise to explicit a definition of $Y_n(x)$. The expression (0.1.3) is especially useful when required to obtain the approximation rate of T_n. Indeed, suppose that the modulus of continuity of f satisfies $\omega_f(n\delta) \leqslant n\omega_f(\delta)$ for any $n \in \mathbb{N}$ and $\delta > 0$. This is the case, for instance, when $\mathbb{M}$ is a smooth compact submanifold of Euclidean space together with its geodesic distance. Then for any sequence $\delta_n \to 0$ we have that

$$\begin{aligned}\|T_n f - f\|_\infty &\leqslant \omega_f(\delta_n)\left\{2 + \delta_n^{-1}\sup_{x\in\mathbb{M}}\mathbb{E}\left[d(Y_n(x), x)\mathbb{1}_{d(Y_n(x),x)\geqslant\delta_n}\right]\right\} \\ &\leqslant \omega_f(\delta_n)\sup_{x\in\mathbb{M}}\left\{2 + \delta_n^{-2}\sup_{x\in\mathbb{M}}\mathbb{E}\left[d(Y_n(x), x)^2\right]\right\}.\end{aligned}$$

We may show that $\sup_{x\in\mathbb{M}}\delta_n^{-1}\mathbb{E}[d(Y_n(x), x)]$ is uniformly bounded in n, which therefore entails that

$$\|T_n f - f\|_\infty = \mathcal{O}(\omega_f(\delta_n)).$$

The square of the distance function is easier to deal with in some cases, such as when $d(x, y) = |x - y|$ on $[0, 1]$. In this case, if $\sup_{x\in\mathbb{M}}\mathbb{E}[d(Y_n(x), x)^2] \leqslant \sigma_n^2$ for some sequence of "variances" $\sigma_n^2 \in \mathbb{R}$, then by letting $\delta_n = \sigma_n^{-1}$ we obtain that

$$\|T_n f - f\|_\infty = \mathcal{O}(\omega_f(\sigma_n^{-1})).$$

This relates the uniform contraction rate of the variables $Y_n(x)$ around x to the approximation rate of T_n.

Interpretability and shape constrained estimation. It is notoriously difficult to elicit priors on infinite dimensional spaces. The use of a sieve priors such as (0.1.2) reduces the problem to that of eliciting a prior on the finite dimensional subsets $T_n(\mathbb{F})$, which always admit a representation of the form

$$T_n(\mathbb{F}) = \left\{ \sum_{j=0}^{d_n} c_{j,n}\, \phi_{j,n} \right\}$$

for some basis densities $\phi_{j,n}$ and coefficients $c_{j,n}$, and a prior ρ on the parameter n. This parameter n may be thought as representing the complexity of the sieve through its dimension d_n. The asymptotic theory of Chapter 3 suggests taking $\rho(n) \asymp -d_n \log(d_n)$ or $\rho(n) \asymp -d_n$, and also provides some guidance for the choice of the prior on $T_n(\mathbb{F})$. The Bayes estimator resulting from (0.1.2) is simply the mixture of the Bayes estimator obtained from the priors Π_n on $T_n(\mathbb{F})$, weighted by the posterior probabilities of each model.

In some cases, the operators T_n may be extended to act upon probability measures; see again Chapter 3 for more details. If $\mathcal{D}$ is a Dirichlet Process, then the prior induced by the random density $T_N(\mathcal{D})$ where $N \sim \rho$ is independent of $\mathcal{D}$ is both a Dirichlet Process Mixture and a sieve prior as in (0.1.2). The interpretation as a Dirichlet Process Mixture is especially useful in view of the computational methods developed in Kalli et al. (2011). The sieve prior representation is otherwise typically more suited to reversible jump MCMC algorithms for posterior simulation.

We may also want to incorporate very precise types of prior information into the model. For instance, if f_0 is defined on $\mathbb{M} = [0,1]^2$, we may know a priori its marginal distributions. Or we may know that f_0 defined on $[0,1]$ is monotonous.

The sieve prior (0.1.2) is particularly well suited to the incorporation of such shape constraints.

Indeed, it suffices restrict $\mathbb{F}$ to be the set of all bounded densities satisfying the required shape constraint, and to let T_n be such that $T_n(\mathbb{F}) \subset \mathbb{F}$. This is possible in mentioned particular cases, e.g. for copula density estimation and monotone density estimation. The theory continues to apply in this context with still the same interpretability and with rates of convergences depending on the dimensions of $T_n(\mathbb{F})$.

0.1.4 Circular statistics

The above theory has been developed concurrently to the study of a circular analogue to the Bernstein polynomial densities which we use in Chapter 3 to construct sieve priors on circular density spaces. The density basis of the trigonometric polynomials that we consider is given by

$$C_{j,n}(u) \propto \left(1 + \cos\left(u - \tfrac{2\pi j}{2n+1}\right)\right)^n$$

with a known normalizing constant and $j \in \{0, 1, 2, \ldots, 2n\}$. The central element $C_{0,n}$ is, up to a multiplicative constant, the De la Vallée Poussin kernel studied in Pólya and Schoenberg (1958). The consideration of this set of translates was proposed in Róth et al. (2009) in the context of Computer Aided Geometric Design. Here we have studied properties of $C_{j,n}$ which are particularly relevant to mixture modelling.

As such, we provide the Fourier coefficients of $C_{j,n}$ which are also referred to in the directional statistics literature as the trigonometric moments. These provide the change of basis formula between the $C_{j,n}$ and the usual trigonometric basis $\{1, \cos(u), \sin(u), \ldots, \cos(nu), \sin(nu)\}$. Together with the method for the efficient simulation of the $C_{j,n}$ provided in Chapter 3 and the characterization of

positive trigonometric densities as mixtures of the $C_{j,n}$, this shows how any positive trigonometric density can be directly simulated as a mixture and provides an algorithm to do so.

Some properties of a mixture density $f = \sum_{j=0}^{2n} c_{j,n} C_{j,n}$, where $c_{j,n} \geqslant 0$, $\sum_j c_{j,n} = 1$, can also be easily related to properties of the vector of coefficients $(c_{j,n})_{j=0}^{2n}$. Those can be neatly stated in terms of properties of the operator

$$T_n f = \sum_{j=0}^{2n} \int_{R_{j,n}} f(u)du \, C_{j,n} \tag{0.1.4}$$

with $R_{j,n} = \left[\frac{\pi(2j-1)}{2n+1}, \frac{\pi(2j+1)}{2n+1}\right)$. Using variation diminishing properties of the De la Vallée Poussin kernel studied in Pólya and Schoenberg (1958), it is shown in Chapter 3 that T_n reproduces constants, that it preserves periodic unimodality and diminishes total variation. Furthermore, $\|T_n f - f\|_\infty \to 0$ as $n \to \infty$ for every continuous f. The same kind of properties hold for the De la Vallée Poussin means

$$V_n f(x) = \int_0^{2\pi} f(x-u) C_{0,n}(u) du$$

for which it is also known that $\|V_n f - f\|_\infty = \mathcal{O}(\omega_f(n^{-1/2}))$ (Lorentz, 1986). Approximation rates can be obtained for T_n defined in (0.1.4) using the technique described in Section 3.4.4.

In order to showcase the practical usefulness of these densities and of the framework which we used to construct sieve priors, we have compared the finite sample performance of our Bayes estimators to other circular density estimators based on trigonometric polynomial densities. Notably, Fernández-Durán (2004) used a surjective parameterization of the space of all trigonometric densities through a complex hypersphere in order to compute maximum likelihood estimators. Model dimensions are chosen using the AIC or BIC criteria. In Fernández-Durán and Gregorio-Domínguez (2016a), posterior means are also considered. The density

estimators based on our models provide the best performance in a variety of scenarios. These results are not meant to show that our estimators are best-possible, but simply that the De la Vallée Poussin basis should be considered for circular density modelling, especially when there is an availability of prior information to support informed Bayesian estimation.

CHAPTER I

METRICS AND DIVERGENCES FOR PROBABILITY MEASURES

Let $(\mathbb{M}, d)$ be a complete and separable metric space together with its Borel σ-algebra $\mathfrak{B}_{\mathbb{M}}$ and let $\mathcal{M}$ be the space of all probability measures on $(\mathbb{M}, \mathfrak{B}_{\mathbb{M}})$. This section presents elementary facts about $\mathcal{M}$ and its common metrics and topologies, some of which may be found in Aliprantis and Border (2006); Ghosh and Ramamoorthi (2003a); Gibbs and Su (2002); Billingsley (2013).

1.1 The total variation distance

The space $\mathcal{M}$ embeds in the (complete) normed linear space of measures μ with finite total variation $\|\mu\|_{\mathrm{TV}} = \sup_{A \in \mathfrak{B}_{\mathbb{M}}} |\mu(A)|$ and inherits the total variation distance $d_{\mathrm{TV}}(\mu, \nu) = \|\mu - \nu\|_{\mathrm{TV}}$. While this metric is easily interpretable as measuring worst case difference in mass allocation, it is so at the loss of tractability of the resulting metric space: $\mathcal{M}$, with the total variation distance, is not separable unless $\mathbb{M}$ is countable.[1]

However, the problem disappears when considering dominated subsets of $\mathcal{M}$ as such sets identify with part of a suitable L^1 space.

Lemma 1.1.1. *A subset $\mathcal{F} \subset \mathcal{M}$ is dominated by a σ-finite measure if and only if $(\mathcal{F}, d_{TV})$ is separable.*

Proof. First suppose $\mathcal{F}$ is dominated by a σ-finite measure λ. Take $\mu, \nu \in \mathcal{F}$ and consider the densities (i.e. Radon-Lebesgue-Nikodym derivatives) $f = d\mu/d\lambda$,

[1] To see this non-separability, suppose $\mathbb{M}$ is uncountable and consider the subset $\{\delta_x\}_{x \in \mathbb{M}}$ of point mass measures. Let also $E \subset \mathcal{M}$ be such that for every $x \in \mathbb{M}$, there exists $\nu_x \in E$ with $\|\delta_x - \nu_x\|_{\mathrm{TV}} < 1/2$. It follows that ν_x must contain a point mass at x. Since ν_x is finite, it contains only a countable number of such point masses. Any countable number of such measures can only approximate in this way a countable subset of $\{\delta_x\}_{x \in \mathbb{M}}$. This shows E is uncountable and hence $\mathcal{M}$ is not separable.

$g = d\nu/d\lambda$ in $L^1(\lambda)$. The set $A = \{x \in \mathbb{M} \mid f(x) \geqslant g(x)\} \in \mathfrak{B}_{\mathbb{M}}$ is such that $\|\mu - \nu\|_{\mathrm{TV}} = (\mu - \nu)(A) = (\nu - \mu)(\mathbb{M}\backslash A)$ and it follows that

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2}\int |f - g|\, d\lambda. \tag{1.1.1}$$

Hence d_{TV} is equivalent to the L^1 distance on the identification of $\mathcal{F}$ with its densities $\{d\mu/d\lambda \mid \mu \in \mathcal{F}\} \subset L^1(\lambda)$. Since λ is σ-finite and $\mathfrak{B}_{\mathbb{M}}$ countably generated, $L^1(\lambda)$ is separable and so must be $(\mathcal{F}, d_{\mathrm{TV}})$.

Conversely, if $(\mathcal{F}, d_{\mathrm{TV}})$ is separable, let $E = \{\mu_n \mid n \in \mathbb{N}\}$ be a countable dense subset of $\mathcal{F}$. We show that $\lambda = \sum_{n\in\mathbb{N}} \mu_n 2^{-n}$ dominates $\mathcal{F}$. Let $A \in \mathfrak{B}_{\mathbb{M}}$ be such that $\lambda(A) = 0$, fix $\mu \in \mathcal{F}$ and $\varepsilon > 0$. Then $\mu_n(A) = 0$ for every n and by density of E there exists a $n \in \mathbb{N}$ such that $\mu(A) = |\mu(A) - \mu_n(A)| < \varepsilon$. Since $\varepsilon > 0$ was arbitrary, $\mu(A) = 0$. This shows $\mu \ll \lambda$ for every $\mu \in \mathcal{F}$. □

1.2 The Prokhorov metric and weak convergence

As to obtain a complete separable metric structure on $\mathcal{M}$ we may relax the total variation distance the Prokhorov metric d_P. It is defined as

$$d_P(\mu, \nu) = \inf\{\varepsilon > 0 \mid \forall A \in \mathfrak{B}_{\mathbb{M}},\ \mu(A) \leqslant \nu(A^\varepsilon) + \varepsilon\} \tag{1.2.1}$$

where $A^\varepsilon = \{x \in \mathbb{M} \mid d(x, A) < \varepsilon\}$ is the ε-neighborhood of A and $d(x, A) = \inf\{d(x, y) \mid y \in A\}$ (Strassen, 1965; Prokhorov, 1956). This provides a metrization of the topology of weak convergence of probability measures (see for isntance Huber (2011)), also known as the weak-* topology of the continuous dual of $C_b(\mathbb{M})$, which is further described by the Portmanteau theorem (see Billingsley (2013)).

It follows from (1.2.1) that $d_P \leqslant d_{\mathrm{TV}}$. Hence total variation convergence implies weak convergence. The converse obviously does not hold, as can be seen by considering the sequence of measures $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{i/n}$ defined on $[0, 1] \subset \mathbb{R}$.

While $\{\mu_n\}$ does not converge in $(\mathcal{M}, d_{\mathrm{TV}})$, it converges to the Lebesgue measure in $(\mathcal{M}, d_P)$. The approximation properties of measures with finite support are further discussed in the proof of the following lemma.

Lemma 1.2.1. *The space $(\mathcal{M}, d_P)$ is separable if and only if $(\mathbb{M}, d)$ is separable.*

Proof. Suppose $(\mathcal{M}, d_P)$ is separable and consider the map $\phi : \mathbb{M} \to \mathcal{M} : x \mapsto \delta_x$. Fix $x, y \in \mathbb{M}$ and let $\varepsilon = \min\{d(x,y), 1\}$. The fact that $\delta_x(\{x\}) \leqslant \delta_y(\{A\}^\varepsilon) + \varepsilon$ shows $d_P(\phi(x), \phi(y)) \geqslant \varepsilon$ and obviously $\delta_x(A) \leqslant \delta_y(A^\delta) + \delta$ for every $\delta > \varepsilon$ and $A \in \mathfrak{B}_{\mathbb{M}}$. Hence $d_P(\phi(x), \phi(y)) = \min\{d(x,y), 1\}$ and ϕ establishes an isometry between $(\mathbb{M}, \tilde{d})$ and $(\mathcal{M}, d_P)$ when $\tilde{d} = \min\{d, 1\}$. Thus $(\mathbb{M}, \tilde{d})$ is separable and so is the homeomorphic $(\mathbb{M}, d)$.

Now suppose $(\mathbb{M}, d)$ is separable and let E be a countable dense subset. We show that $\{\sum_{i=1}^n \alpha_i \delta_{x_i} \mid n \in \mathbb{N},\, \alpha_i \in \mathbb{Q},\, x_i \in E\}$ is dense in $(\mathcal{M}, d_P)$. To this end, fix $\varepsilon > 0$ and let $\mu \in \mathcal{M}$. Let $n \in \mathbb{N}$ and $\{x_i\}_{i=1}^n \subset E$ be such that $\mu\left(\mathbb{M} \backslash \bigcup_{i=1}^n B(x_i, \varepsilon)\right) < \varepsilon/2$ and consider $\nu = \sum_{i=1}^n \alpha_i \delta_{x_i}$ where $\alpha_i \in \mathbb{Q}$ is such that $|\alpha_i - \mu(B(x_i, \varepsilon))| < \varepsilon/(2n)$. Then for any $A \in \mathfrak{B}_{\mathbb{M}}$,

$$\mu(A) \leqslant \frac{\varepsilon}{2} + \sum_{i=1}^n \mu(B(x_i, \varepsilon/2) \cap A) \leqslant \varepsilon + \sum_{x_i \in A^\varepsilon} \alpha_i = \nu(A^\varepsilon) + \varepsilon$$

which shows $d_P(\mu, \nu) \leqslant \varepsilon$. □

The following theorem provides *coupling* characterizations of the two metrics seen thus far and highlights how exactly d_P weakens d_{TV} by taking into account the metric structure of $\mathbb{M}$. A proof can be found in Dudley (2002) and here we denote $A^{\delta]} = \{x \in \mathbb{M} \mid d(x, A) \leqslant \delta\}$.

Theorem 1.2.2 (Strassen (1965)). *Let $\mu, \nu \in \mathcal{M}$ and let $\mathcal{C}$ be the set of all pairs (X, Y) of random variables on $(\mathbb{M}, \mathfrak{B}_{\mathbb{M}})$ (defined on some common probability*

space of probability measure $\mathbb{P}$) with marginal distributions μ and ν, respectively. Then for every $\varepsilon, \delta \geqslant 0$ following two statements are equivalent:

(i) for every $A \in \mathfrak{B}_{\mathbb{M}}$, $\mu(A) \leqslant \nu(A^{\delta]}) + \varepsilon$;

(ii) there exists $(X, Y) \in \mathcal{C}$ such that $\mathbb{P}(d(X, Y) > \delta) \leqslant \varepsilon$.

Considering the cases $\delta = 0$ and $\delta = \varepsilon$ yields explicit descriptions of d_{TV} and d_P.

Corollary 1.2.3. *Let μ, ν and $\mathcal{C}$ be as in Theorem 1.2.2. Then*

$$d_{TV}(\mu, \nu) = \inf_{(X,Y)\in\mathcal{C}} \{\varepsilon > 0 \mid \mathbb{P}(d(X, Y) > 0) \leqslant \varepsilon\}$$

and

$$d_P(\mu, \nu) = \inf_{(X,Y)\in\mathcal{C}} \{\varepsilon > 0 \mid \mathbb{P}(d(X, Y) > \varepsilon) \leqslant \varepsilon\}.$$

We conclude the presentation of $(\mathcal{M}, d_P)$ with a simple measurability result relevant to the definition of random probability measures.

Lemma 1.2.4. *The evaluation maps $\mathcal{M} \ni \mu \mapsto \mu(A)$, where $A \in \mathfrak{B}_{\mathbb{M}}$, are Borel measurable.*

Proof. The definition of d_P entails the map $f(\mu) = \mu(A)$ is upper semi-continuous whenever A is a closed set. Indeed, fix $\mu_0 \in \mathcal{M}$, $\varepsilon > 0$ and let $\delta > 0$ be such that $\delta < \varepsilon/2$ and $\mu_0(A^{\delta} \backslash A) < \varepsilon/2$. Then by definition $d_P(\mu, \mu_0) < \delta$ implies $\mu(A) - \mu_0(A) \leqslant \mu_0(A^{\delta} \backslash A) + \delta < \varepsilon$. Since semi-continuous functions are measurable, this shows the family $\mathcal{A}$ of sets $A \in \mathfrak{B}_{\mathbb{M}}$ such that $\mu \mapsto \mu(A)$ is measurable contains the π-system of closed sets of $\mathbb{M}$. It is immediate to verify $\mathcal{A}$ is also a λ-system. From Dynkin's theorem, we obtain that $\mathcal{A} \supset \mathfrak{B}_{\mathbb{M}}$. □

1.3 The Kullback-Leibler divergence

We now turn to another measure of discrepancy between probability measures, introduced by Kullback and Leibler (1951).

To motivate its definition, let $\lambda, \mu, \nu \in \mathcal{M}$ and consider the problem of testing $H_1 : \lambda = \mu$ versus $H_2 : \lambda = \nu$ given an i.i.d. sample $\{X_i\}_{i=1}^n$ of size $n \in \mathbb{N}$ with common distribution λ. We write $X = (X_1, \ldots, X_n) \sim \lambda^{(n)}$. Assuming μ, ν and λ are mutually absolutely continuous[2], we can define their likelihood ratio as

$$\frac{d\mu}{d\nu}(X) := \prod_{i=1}^{n} \frac{d\mu}{d\nu}(X_i) \in [0, \infty]. \tag{1.3.1}$$

The *weight of evidence* brought by the sample X in favor of H_1 versus H_2 is defined as $W(X) = \log \frac{d\mu}{d\nu}(X)$ (Good, 1985). This is also known as the relative information of X according to (μ, ν) (Sason and Verdú, 2016), and is the usual statistic of the likelihood ratio test.

The Kullback-Leibler divergence $D(\mu\|\nu)$ between μ and ν is the expected weight of evidence brought by a single observation taken under the hypothesis H_1 (taking the expectation under H_2 simply reverses the sign). In the words of Kullback and Leibler, it is the "mean information for discrimination between H_1 and H_2 per observation". Hence formally

$$D(\mu\|\nu) = \int_{\mathbb{M}} \log\left(\frac{d\mu}{d\nu}\right) d\mu. \tag{1.3.2}$$

When μ, ν are not absolutely continuous with respect to one another, we define $D(\mu\|\nu) = \infty$. This case does not always require particular care as we may introduce $\xi = \mu + \nu$ and write $D(\mu\|\nu) = \int_{f>0} f \log(f/g) d\xi$ with $f = d\mu/d\xi$ and

[2]This assumption is not completely necessary, but it is enforced here as to simplify the discussion.

$g = d\nu/d\xi$. Equation (1.3.2) is well defined as the integral of the negative part of the integrand is bounded:

$$\begin{aligned} -\int_{\{d\mu/d\nu<1\}} \log\left(\frac{d\mu}{d\nu}\right) d\mu &\leqslant \int_{\{d\mu/d\nu<1\}} \left(\frac{d\nu}{d\mu} - 1\right) d\mu \\ &= \nu(\{d\mu/d\nu < 1\}) - \mu(\{d\mu/d\nu < 1\}) \\ &< 1. \end{aligned}$$

The following theorem shows that the magnitude of D is a statistically meaningful quantity. Here we let $\mathbb{Y}$ be some measurable space and $T : \mathbb{M} \to \mathbb{M}$ a measurable transform. We denote by μT^{-1} the pushforward measure defined by $\mu T^{-1}(A) = \mu(T^{-1}(A))$ for $A \subset \mathbb{Y}$ measurable.

Theorem 1.3.1 (Kullback and Leibler). *If $\mu, \nu \in \mathcal{M}$ are absolutely continuous with respect to one another and if $T : \mathbb{M} \to \mathbb{Y}$ is a measurable transform of $\mathbb{M}$, then $D(\mu T^{-1} \| \nu T^{-1}) \leqslant D(\mu\|\nu)$ with equality if and only if T is a sufficient statistic for $\{\mu, \nu\}$.*

Proof. To ease notation, write $\tilde{\mu} = \mu T^{-1}$, $\tilde{\nu} = \nu T^{-1}$ and let $h = \frac{d\mu}{d\nu}/(\frac{d\tilde{\mu}}{d\tilde{\nu}} \circ T)$. Remark that by change of variable

$$D(\mu\|\nu) - D(\tilde{\mu}\|\tilde{\nu}) = \int_{\mathbb{M}} \left(\log\left(\tfrac{d\mu}{d\nu}\right) - \log\left(\tfrac{d\tilde{\mu}}{d\tilde{\nu}} \circ T\right)\right) d\mu = \int_{\mathbb{M}} \log(h)\, d\mu.$$

Since $\int_{\mathbb{M}} 1/h\, d\mu = 1$, Jensen's inequality together with the convexity of the function $\phi(t) = t\log(t)$ yields $D(\mu\|\nu) - D(\tilde{\mu}\|\tilde{\nu}) \geqslant \phi(1) = 0$ with equality if and only if $h = 1$ μ-almost surely. Hence equality happens if and only if $\frac{d\mu}{d\nu} = \frac{d\tilde{\mu}}{d\tilde{\nu}} \circ T$ μ-almost everywhere which, since $\mu \ll \nu$ and $\nu \ll \mu$, amounts to saying T is sufficient for $\{\mu, \nu\}$ (Halmos and Savage, 1949, Theorem 1). □

Considering the case where T is constant yields an equally important result.

Corollary 1.3.2. *If $\mu, \nu \in \mathcal{M}$, then $D(\mu\|\nu) \geqslant 0$ with equality if and only if $\mu = \nu$.*

1.3.1 Exponential convergence of the likelihood ratio

The importance of the Kullback-Leibler divergence in Bayesian nonparametrics and asymptotic statistics stems from its characterization of the likelihood ratio's exponential convergence.

Proposition 1.3.3. *Suppose $\mu, \nu \in \mathcal{M}$ are absolutely continuous with respect to one another and let $\{X_i \mid i \in \mathbb{N}\}$ contain independent random variables with distribution μ. The following two statements are equivalent.*

(i) $D(\mu\|\nu) < \infty$.

(ii) There exists an $R \in [0, \infty)$ such that $\prod_{i=1}^n \frac{d\mu}{d\nu}(X_i) = \exp\{nR + o(n)\}$ almost surely.

Also, when the statements hold, we have $R = D(\mu\|\nu)$ in (ii).

Proof. Note that *(ii)* is equivalent to $\frac{1}{n}\sum_{i=1}^n \log \frac{d\mu}{d\nu}(X_i) = R + o(1)$ almost surely for some $R \in [0, \infty)$. By the Strong law of large numbers, this happens if and only if $R = \mathbb{E}\left[\log \frac{d\mu}{d\nu}(X_i)\right] = D(\mu\|\nu)$. □

Given a bound on the second moment of $d\mu/d\nu$ also provides a stochastic control on the fluctuations of the likelihood ratio. Here we state a result of this kind which is particularly helpful to the study of posterior distribution.

Lemma 1.3.4 (Lemma 8.1 of Ghosal et al. (2000))**.** *Let Π be a prior on a subset $\mathbb{F}$ of $(\mathcal{M}, d_{TV})$ with respect to its Borel σ-algebra. Fix $\varepsilon > 0$, $\delta_1 > 0$, $\delta_2 > 0$ and let $W = \{\mu \in \mathbb{F} \mid \int \log \frac{d\nu}{d\mu} d\nu \leqslant \delta_1, \int \left(\log \frac{d\nu}{d\mu}\right)^2 d\nu \leqslant \delta_2\}$. If $(X_i)_{i=1}^n$ is a sequence of independent random variables with distribution ν, then*

$$\int_W \prod_{i=1}^n \frac{d\mu}{d\nu}(X_i)\Pi(d\mu) \geqslant e^{-n(\delta_1+\varepsilon)}\Pi(W) \tag{1.3.3}$$

holds with probability at least $1 - \frac{\delta_2}{n\varepsilon^2}$.

Remark 1.3.1. The Lemma shows how the Kullback-Leibler divergence, here controlled through the constant δ_1, provides a probable exponential bound on the convergence of (an average) of the likelihood ratio. The second moment bound δ_2 of the log likelihood ratio acts linearly on the probability of the lower bound.

Proof of Lemma 1.3.4. Assume, without loss of generality, that $\Pi(W) > 0$ and let $\tilde{\Pi} = \Pi/\Pi(W)$ be the renormalization of Π over W. By Jensen's inequality,

$$\log \int_W \prod_{i=1}^n \frac{d\mu}{d\nu}(X_i)\tilde{\Pi}(d\mu) \geqslant \sum_{i=1}^n \int_W \log \frac{d\mu}{d\nu}(X_i)\tilde{\Pi}(d\mu)$$

and hence the complementary probability of (1.3.3) is

$$\mathbb{P}\left(\int_W \prod_{i=1}^n \frac{d\mu}{d\nu}(X_i)\tilde{\Pi}(d\mu) \leqslant e^{-n(\delta_1+\varepsilon)}\right) \leqslant \mathbb{P}\left(\sum_{i=1}^n \int_W \log \frac{d\mu}{d\nu}(X_i)\tilde{\Pi}(d\mu) \leqslant -n(\delta_1+\varepsilon)\right).$$

Subtracting the average $E = -n\int_W D(\nu\|\mu)\tilde{\Pi}(d\mu)$ of $\sum_{i=1}^n \int_W \prod_{i=1}^n \frac{d\mu}{d\nu}(X_i)\tilde{\Pi}(d\mu)$ and using the fact that $E \geqslant -n\delta_1$, this is upper bounded by

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^n \int_W \log \frac{d\mu}{d\nu}(X_i)\tilde{\Pi}(d\mu) - E \leqslant -n\varepsilon\right) \\
&\leqslant \mathbb{P}\left(\left(\sum_{i=1}^n \int_W \log \frac{d\mu}{d\nu}(X_i)\tilde{\Pi}(d\mu) - E\right)^2 \leqslant (n\varepsilon)^2\right) \\
&\leqslant \frac{1}{n\varepsilon^2}\mathbb{E}\left[\left(\int_W \log \frac{d\mu}{d\nu}(X_i)\tilde{\Pi}(d\mu)\right)^2\right]
\end{aligned}$$

where the last inequality follows by Chebychev's inequality. By Jensen's inequality, $\left(\int_W \log \frac{d\mu}{d\nu}(X_i)\tilde{\Pi}(d\mu)\right)^2 \leqslant \int_W \left(\log \frac{d\mu}{d\nu}(X_i)\right)^2 \tilde{\Pi}(d\mu)$ and by Fubini's theorem and the definition of W we obtain that the expectation of $\int_W \left(\log \frac{d\mu}{d\nu}(X_i)\right)^2 \tilde{\Pi}(d\mu)$ is bounded by δ_2. □

1.4 f-divergences

A large and very useful family of measures of discrepancy between probability measures is obtained by considering expected transforms of the likelihood ratio. The basic idea is that, for the purposes of likelihood based inference, any meaningful measure of distance or discrepancy between probability measures should be function of their likelihood ratio. In particular, expected convex transforms of the likelihood ratio encompass a number of useful particular cases and share useful properties.

Definition 1.4.1. Let $f : [0, \infty] \to (-\infty, \infty]$ be a convex function which is strictly convex at 1 and such that $f(1) = 0$. Given two probability measures $\mu, \nu \in \mathbb{M}$ such that $\mu \ll \nu$, the f-divergence between μ and ν is defined as

$$D_f(\mu \| \nu) = \mathbb{E}_\nu \left[f\left(\frac{d\mu}{d\nu}\right)\right] = \int f\left(\frac{d\mu}{d\nu}\right) d\nu. \tag{1.4.1}$$

Remark 1.4.1. Part *(i)* of Proposition 1.4.2 shows that D_f is well-defined: while it may be infinite, the integral of the negative part of $f(d\mu/d\nu)$ with respect to ν is always finite.

Table 1.1 summarizes a few of the most common f-divergences.

Remark 1.4.2 (Hilbert interpretation). Let $\mathbb{F}$ be a separable subset of $(\mathcal{M}, d_{\mathrm{TV}})$ and let λ be a dominating σ-finite measure. While $\mathbb{F}$ is naturally identifiable with part of $L^1(\lambda)$ (see the proof of Lemma 1.1.1), the identification $\mu \mapsto \sqrt{d\mu/d\lambda} \in L^2(\lambda)$ with part of the unit sphere of the Hilbert space $L^2(\lambda)$ provides additional tools. The resulting inner product of L^2 is referred to as the (1/2)-affinity defined by

$$A_{1/2}(\mu, \nu) = \left\langle \sqrt{d\mu/d\lambda}, \sqrt{d\nu/d\lambda} \right\rangle_{L^2(\lambda)}.$$

and the $L^2(\lambda)$-distance for root densities is referred to as the Hellinger distance

$$H(\mu, \nu) = \| \sqrt{d\mu/d\lambda} - \sqrt{d\nu/d\lambda} \|_{L^2(\lambda)}.$$

Table 1.1: Common f-divergence definitions and related functions.

f-divergence	Symbol	$f(t)$
Total variation distance	$d_{\mathrm{TV}}(\mu,\nu)$	$\frac{1}{2}\lvert t-1\rvert$, $\max\{0,t-1\}$
E_γ divergence	$E_\gamma(\mu,\nu)$	$\max\{0,t-\gamma\}$
Kullback-Leibler divergence	$D(\mu\Vert\nu)$	$t\log(t)$
Squared Hellinger distance	$H(\mu,\nu)^2$	$(\sqrt{t}-1)^2$
χ^2-divergence	$\chi^2(\mu,\nu)$	(t^2-1), $(t-1)^2$
Hellinger divergence of order $\alpha>0$	$\mathcal{H}_\alpha(\mu,\nu)$	$(t^\alpha-1)/(\alpha-1)$
Related functions		
Hellinger distance	$H(\mu,\nu)=\left(\int\left(\sqrt{d\mu/d\nu}-1\right)^2 d\nu\right)^{1/2}$	
α-affinity ($\alpha>0$)	$A_\alpha(\mu,\nu)=\mathbb{E}_\nu\left[(d\mu/d\nu)^\alpha\right]$	
Rényi divergence of order $\alpha>0$	$D_\alpha(\mu,\nu)=\log\left(A_\alpha(\mu,\nu)\right)/(\alpha-1)$	

These quantities are monotonous transforms of the Hellinger divergence, of the Rényi divergence and of α-affinity.

Proposition 1.4.2. *Let D_f be any f-divergence, as in Definition 1.4.1.*

(i) $\mathbb{E}_\nu\left[\max\{0,-f\left(d\mu/d\nu\right)\}\right]<\infty$

(ii) We have $D_f(\mu\Vert\nu)\geqslant 0$ with $D_f(\mu\Vert\nu)=0$ if and only if $\mu=\nu$.

(iii) If T is any measurable transform of $\mathbb{M}$, then $D_f(\mu\Vert\nu)\geqslant D_f(\mu T^{-1},\nu T^{-1})$ with equality if and only if T is a sufficient statistic for $\{\mu,\nu\}$.

Proof. (i) Since f is convex with $f(1)=0$, either $f(t)\geqslant 0$ for every $t\geqslant 1$ or

$f(t) \geqslant 0$ for every $0 \leqslant t \leqslant 1$. In the first case,

$$\begin{aligned}\mathbb{E}_\nu\left[\max\{0, -f\left(d\mu/d\nu\right)\}\right] &\leqslant \mathbb{E}_\nu\left[-f\left(\max\{d\mu/d\nu, 1\}\right)\right] \\ &\leqslant -f\left(\mathbb{E}_\nu\left[\max\{d\mu/d\nu, 1\}\right]\right) < \infty.\end{aligned}$$

by Jensen's inequality. The second case follows similarily.

Taking T any constant function, *(ii)* is seen to be a particular case of *(iii)*.

In order to prove *(iii)*, let $\tilde{\mu} = \mu T^{-1}$, $\tilde{\nu} = \nu T^{-1}$ and note that

$$\frac{d\tilde{\mu}}{d\tilde{\nu}}(y) = \mathbb{E}_{X\sim\nu}\left[\frac{d\mu}{d\nu}(X) \mid T(X) = y\right].$$

Hence using Jensen's inequality and with $X \sim \nu$ we find

$$\begin{aligned}D_f(\mu\|\nu) &= \mathbb{E}\left[f\left(\frac{d\mu}{d\nu}(X)\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[f\left(\frac{d\mu}{d\nu}(X)\right) \mid T(X)\right]\right] \\ &\geqslant \mathbb{E}\left[f\left(\mathbb{E}\left[\frac{d\mu}{d\nu}(X) \mid T(X)\right]\right)\right] \\ &= \mathbb{E}\left[f\left(\frac{d\tilde{\mu}}{d\tilde{\nu}}(T(X))\right)\right] = D_f(\tilde{\mu}\|\tilde{\nu}).\end{aligned}$$

Since f is strictly convex at 1, equality holds if and only if $\frac{d\mu}{d\nu} = \frac{d\tilde{\mu}}{d\tilde{\nu}} \circ T$ ν-almost everywhere. Because $\mu \ll \nu$, this is the same as saying that T is a sufficient statistic for $\{\mu, \nu\}$ (Halmos and Savage, 1949, Theorem 1). □

1.4.1 Application in importance sampling

One particularly accessible and useful subject in which f-divergences appear is in error quantification for importance sampling. Without delving very deep in the theory (see Chatterjee and Diaconis (2018); Agapiou et al. (2017); Sanz-Alonso (2018) for more details and arguably converse results on necessary sample sizes),

let me introduce the problem and state expected error bounds in terms of f-divergences.

Let φ be integrable with respect to a measure ν and define $I(\varphi) = \int \varphi \, d\nu$. The goal is to estimate $I(\varphi)$ using a sample $\{X_i\}_{i=1}^n$ of independent random variables with identical distribution μ satisfying $\nu \ll \mu$. To this end, let

$$I_n(\varphi; \mu) = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i) \frac{d\nu}{d\mu}(X_i)$$

and notice that $\mathbb{E}\left[I_n(\varphi; \mu)\right] = I(\varphi)$. The law of large number entails almost sure convergence of $I_n(\varphi; \mu)$ to $I(\varphi)$, and the almost sure convergence rate $|I_n(\varphi; \mu - I(\varphi))| = o\left(n^{1/p-1}\right)$ is provided by the MZ Theorem under the assumption $\|\varphi \, d\nu/d\mu\|_{L^p(\mu)} < \infty$ for some $1 \leqslant p < 2$. If $\|\varphi \, d\nu/d\mu\|_{L^2(\mu)} < \infty$, then the Central Limit Theorem yields confidence intervals.

For the study of expected errors, it is an easy exercise to see that the variance of $I_n(\varphi; \mu)$ is minimized at $I_n(\varphi; \mu^*)$, where μ^* is such that $d\mu^*/d\nu = |\varphi|/I(|\varphi|)$. In this case, $\mathrm{Var}(I_n(\varphi; \mu^*)) = (I(|\varphi|)^2 - I(\varphi)^2)/n$. In general, a straightforward calculation shows that we have

$$\begin{aligned} \mathrm{Var}(I_n(\varphi; \mu)) &= \frac{I(|\varphi|)^2}{n} \chi^2(\mu^*, \mu) + \mathrm{Var}(I_n(\varphi; \mu^*)) \\ &\leqslant I(|\varphi|)^2 \frac{\chi^2(\mu^*, \mu) + 1}{n}. \end{aligned}$$

While the term $I(|\varphi|)^2$ is typically not precisely known in practice, the inequalities

$$\chi^2(\mu^*, \mu) \leqslant \left\| \frac{d\mu^*}{d\mu} \right\|_{L^\infty(\mu)} d_{\mathrm{TV}}(\mu^*, \mu) \leqslant \left\| \frac{d\mu^*}{d\mu} \right\|_{L^\infty(\mu)} - 1,$$

which can be found in (Binette, 2019), can help control the χ^2 divergence.

In the case where $\chi^2(\mu^*, \mu) = \infty$, and consequently $\mathrm{Var}(I_n(\varphi; \mu)) = \infty$, we can still get first moment bounds on the absolute error in terms of the tail distribution of $d\mu^*/d\mu$. The following proposition is a variation on the first part of Theorem

1.1 of Chatterjee and Diaconis (2018). We chose to express the result in term of the tail of $d\mu^*/d\mu$, instead of the tail of $d\nu/d\mu$, as the former incorporates aspects the function φ. Otherwise, minimizing a divergence between ν and μ may be entirely unrelated to the minimization of the expected error.

Proposition 1.4.3. *Let $Y \sim \mu^*$ and $\rho = d\mu^*/d\mu$. Then for every $a \geqslant 0$,*

$$\mathbb{E}\left[|I_n(\varphi;\mu) - I(\varphi)|\right] \leqslant I(|\varphi|)\left(\sqrt{\frac{a}{n}} + 2\mathbb{P}\left(\rho(Y) > a\right)\right). \tag{1.4.2}$$

Proof. In order to simplify the notation, let $f = I(|\varphi|)\text{sign}(\varphi)$, $h = f\,\mathbb{1}(\rho \leqslant a)$ and define

$$J(f) = \int f\, d\mu^* = I(\varphi)$$

and

$$J_n(f;\mu) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)\rho(X_i) = I_n(\varphi;\mu).$$

Following Chatterjee and Diaconis (2018), write

$$|I_n(\varphi;\mu) - I(\varphi)| \leqslant |J_n(f;\mu) - J_n(h;\mu)| + |J_n(h;\mu) - J(h)| + |J(h) - J(f)|.$$

Now for the first term

$$\begin{aligned}\mathbb{E}\left[|J_n(f;\mu) - J_n(h;\mu)|\right] &\leqslant \mathbb{E}\left[|f(X_1)\rho(X_1) - h(X_1)\rho(X_1)|\right] \\ &= I(|\varphi|)\mathbb{P}\left(\rho(Y) > a\right),\end{aligned}$$

for the third term

$$|J(h) - J(f)| = I(|\varphi|)\mathbb{P}\left(\rho(Y) > a\right),$$

and finally, noting again that $\mathbb{E}\left[J_n(h;\mu)\right] = J(h)$, we find

$$\begin{aligned}\mathbb{E}\left[|J_n(g) - J(h)|\right] &\leqslant \left(\text{Var}(J_n(h))\right)^{1/2} \\ &\leqslant \left(\mathbb{E}\left[(h(X_i)\rho(X_i))^2\right]\right)^{1/2} \\ &\leqslant I(|\varphi|)\sqrt{\frac{a}{n}}.\end{aligned}$$

Combining the above yields the result. □

Controlling the decay rate of the tail probabilities $\mathbb{P}(\rho(Y) > a)$ in terms of Hellinger divergence provides the following bound.

Corollary 1.4.4. *For every $\beta > 0$,*

$$\mathbb{E}\left[|I_n(\varphi;\mu) - I(\varphi)|\right] \leqslant 2I(|\varphi|)\frac{1+\beta\,\mathcal{H}_{\beta+1}(\mu^*,\mu)}{n^{\beta/(1+2\beta)}}.$$

Proof. By Markov's inequality, for any $\beta > 0$,

$$\begin{aligned}\mathbb{P}(\rho(Y) > a) &\leqslant a^{-\beta}\int\left(\frac{d\mu^*}{d\mu}\right)^{\beta} d\mu^* \\ &= a^{-\beta}\int\left(\frac{d\mu^*}{d\mu}\right)^{\beta+1} d\mu \\ &= a^{-\beta}\left(\beta\,\mathcal{H}_{\beta+1}(\mu^*,\mu)+1\right).\end{aligned}$$

With $a = n^{1/(1+2\beta)}$ and combining the above with Proposition 1.4.3 yields the result. □

1.4.2 Application of f-divergences to risk bounds

Cramer-Rao variance bound. Let $\mathbb{F} = \{p_\theta \mid \theta \in \Theta\} \subset L^1(\lambda)$ be a set of densities with respect to the σ-finite measure λ, where $\Theta \subset \mathbb{R}^k$ is open and the map $\theta \mapsto p_\theta$ is injective. We also assume that map $(\theta, x) \mapsto p_\theta(x)$ is sufficiently regular, and we freely interchange integration and differentiation throughout. Now suppose that $\{X_i\}_{i=1}^n$ is a sequence of independent variables with common distribution p_{θ_0} for some $\theta_0 \in \Theta$, and consider an unbiased estimator $\hat{\theta}_n$ of θ_0 which are functions of only $\{X_i\}_{i=1}^n$.

The Kullback-Leibler divergence $\kappa(\theta) := \mathrm{KL}(p_{\theta_0}, p_\theta)$ describes how easily we may discriminate θ_0 from θ, on average, using an observation $X \sim p_{\theta_0}$. Since θ_0 is a minimum of κ, the first order rate of change of κ at this point is zero. The Hessian

of κ, also referred to as Fisher's information matrix, provides more information about variation around θ_0.

Consider, for instance, a direction $u \in \mathbb{R}^k$, $\|u\|_2 = 1$, and let $\kappa''(\theta_0)$ be the second derivative of κ in direction u evaluated at θ_0. Because $\kappa'(\theta_0) = 0$, this is the curvature of κ at θ_0 in direction u. The Cramer-Rao variance bound states that for every $\theta_0 \in \Theta$,

$$\mathbb{E}\left[\langle \hat{\theta}_n - \theta_0, u\rangle^2\right] \geqslant (n\kappa''(\theta_0))^{-1}.$$

That is, the mean squared error of $\hat{\theta}_n$ in direction u is always greater than $1/(n\kappa''(\theta_0))$. Note that the quantity $\kappa''(\theta_0)$ is the same as Fisher's information matrix evaluated as a quadratic form at u.

For the proof, it suffices to consider the case $n = 1$. Let ∂_θ be the differential operator with respect to θ in direction u. For instance, $\partial_{\theta_0}(\log p_{\theta_0})$ is the differential of $\theta \mapsto \log p_\theta$ in direction u evaluated at θ_0. Using the fact that

$$\int \partial_\theta(\log p_\theta) p_\theta \, d\lambda = \int \partial_\theta(p_\theta) \, d\lambda = 0$$

for every $\theta \in \Theta$ and differentiating under the integral, we find that

$$\kappa''(\theta_0) = \mathbb{E}\left[(\partial_{\theta_0}(\log p_{\theta_0}(X)))^2\right].$$

Hence by the Cauchy-Schwartz inequality and integrating by part, we find

$$\begin{aligned}
\sqrt{\mathbb{E}\left[\langle \hat{\theta}_n - \theta_0, u\rangle^2\right] \kappa''(\theta_0)} &\geqslant \int \partial_{\theta_0}(p_{\theta_0})\langle \hat{\theta}_1 - \theta_0, u\rangle \, d\lambda \\
&= \partial_{\theta_0}\left(\int \langle \hat{\theta}_1 - \theta_0, u\rangle p_{\theta_0} \, d\lambda\right) - \int \partial_{\theta_0}(\langle \hat{\theta}_1 - \theta_0, u\rangle) p_{\theta_0} \, d\lambda \\
&= \left\langle \partial_{\theta_0}\left(\int (\hat{\theta}_1 - \theta_0) p_{\theta_0} \, d\lambda\right), u \right\rangle + 1.
\end{aligned}$$

Since $\hat{\theta}_1$ is unbiased, $\int (\hat{\theta}_1 - \theta) p_\theta \, d\lambda = 0$ for every θ, its derivative at θ_0 is also zero, and we obtain the result.

In the biased case, that is if $\mathbb{E}\left[\hat{\theta}_1(X)\right] - \theta_0 = b(\theta_0)$ for some differentiable function b, then a direct adaptation of the above proof yields

$$\mathbb{E}\left[\langle \hat{\theta}_n - \theta_0, u\rangle^2\right] \geqslant \frac{\left(\langle \partial_{\theta_0}(b(\theta_0)), u\rangle + 1\right)^2}{n\kappa''(\theta_0)}.$$

Minimax and Bayes risks lower bounds. Let $\mathbb{F} = \{p_\theta \mid \theta \in \Theta\} \subset L^1(\lambda)$ be a set of densities with respect to the σ-finite measure λ, where Θ is an arbitrary set and the map $\theta \mapsto p_\theta$ is injective. We will also need to assume that $(\theta, x) \mapsto p_\theta(x)$ is measurable in the product space once the Borel σ-algebra of Θ has been introduced. The probability measure corresponding to p_θ is denoted $\mathbb{P}_\theta$ and the expectation under $\mathbb{P}_\theta$ is denoted by $\mathbb{E}_\theta$.

Given a loss function $\ell : \Theta \times \Theta \to [0, \infty)$, we define the minimax estimation risk as

$$R = \inf_{\hat{\theta}} \sup_{\theta_0 \in \Theta} \mathbb{E}_{\theta_0}\left[\ell(\theta_0, \hat{\theta})\right] \tag{1.4.3}$$

where the infimum is taken over all estimators $\hat{\theta}$. It is a lower bound on worst case expected loss. Provided a prior Π on Θ, the Bayes risk associated to Π becomes

$$R_\Pi = \inf_{\hat{\theta}} \int_\Theta \mathbb{E}_{\theta_0}\left[\ell(\theta_0, \hat{\theta})\right] \Pi(d\theta_0). \tag{1.4.4}$$

Note that (1.4.4) is a lower bound on (1.4.3), for any prior Π on Θ.

The Bayes risk can be bounded as follows. Let $B_\varepsilon(\theta_0) = \{\theta \in \Theta \mid \ell(\theta, \theta_0) < \varepsilon\}$ and let $p_{\theta_0,\varepsilon}(x) = \mathbb{E}_{\theta \sim \Pi}\left[p_\theta(x) \mid \ell(\theta, \theta_0) < \varepsilon\right]$ be the density obtained by normalized

averaging over $B_\varepsilon(\theta_0)$, assuming $\Pi(B_\varepsilon(\theta_0)) > 0$. Using these notation, we find

$$\begin{aligned}
R_\Pi &\geqslant \varepsilon \inf_{\hat{\theta}} \int_\Theta \mathbb{P}_\theta \left(\ell(\theta, \hat{\theta}) \geqslant \varepsilon \right) \Pi(d\theta) \\
&\geqslant \varepsilon \left(1 - \int \sup_{\hat{\theta}} \int_\Theta \mathbb{1} \left(\ell(\theta, \hat{\theta}(x)) < \varepsilon \right) p_\theta(x) \Pi(d\theta)\, \lambda(dx) \right) \\
&\geqslant \varepsilon \left(1 - \int \sup_{\theta_0} \int_\Theta \mathbb{1} \left(\ell(\theta, \theta_0) < \varepsilon \right) p_\theta(x) \Pi(d\theta)\, \lambda(dx) \right) \\
&= \varepsilon \left(1 - \int \sup_{\theta_0} \Pi(B_\varepsilon(\theta_0)) p_{\theta_0,\varepsilon}(x)\, \lambda(dx) \right).
\end{aligned}$$

Now let $r_{\Pi,\varepsilon} = 1 - \int \sup_{\theta_0} \Pi(B_\varepsilon(\theta_0)) p_{\theta_0,\varepsilon}(x)\, \lambda(dx)$. Following Theorem II.1 of Guntuboyina (2011), we show that for any f-divergence D_f and any probability measure $Q \ll \lambda$, if $r_{\Pi,\varepsilon} \geqslant 0$ then

$$\int_\Theta D_f(\mathbb{P}_\theta \| Q) \Pi(d\theta) \geqslant W f\left(\frac{1 - r_{\Pi,\varepsilon}}{W} \right) + (1 - W) f\left(\frac{r_{\Pi,\varepsilon}}{1 - W} \right) \tag{1.4.5}$$

where $W = \int \Pi(B_\varepsilon(\tau(x)))\, Q(dx)$ and $\tau(x) = \operatorname{argmax}_{\theta_0} \Pi(B_\varepsilon(\theta_0)) p_{\theta_0,\varepsilon}(x)$. Indeed, with $q = dQ/d\lambda$ and for any $\theta_0 \in \Theta$, we have

$$\begin{aligned}
\mathbb{E}_{\theta\sim\Pi} \left[f\left(\frac{p_\theta}{q} \right) \right] = \Pi(B_\varepsilon(\theta_0)) \mathbb{E}_{\theta\sim\Pi} \left[f\left(\frac{p_\theta}{q} \right) \mid \ell(\theta, \theta_0) < \varepsilon \right] \\
+ \Pi(B_\varepsilon(\theta_0)^c) \mathbb{E}_{\theta\sim\Pi} \left[f\left(\frac{p_\theta}{q} \right) \mid \ell(\theta, \theta_0) \geqslant \varepsilon \right].
\end{aligned}$$

Denoting $p_{\theta_0,\varepsilon^c} = \mathbb{E}_{\theta\sim\Pi} [p_\theta \mid \ell(\theta, \theta_0) \geqslant \varepsilon]$, this is bounded by

$$\Pi(B_\varepsilon(\theta_0)) f\left(\frac{p_{\theta,\varepsilon}}{q} \right) + \Pi(B_\varepsilon(\theta_0)^c) f\left(\frac{p_{\theta,\varepsilon^c}}{q} \right).$$

With $\theta_0 = \tau$ and integrating with respect to Q, we find

$$\begin{aligned}
\int_\Theta D_f(\mathbb{P}_\theta \| Q) \Pi(d\theta) \geqslant \int \Pi(B_\varepsilon(\tau(x))) f\left(\frac{p_{\tau(x),\varepsilon}(x)}{q(x)} \right) Q(dx) \\
+ \int \Pi(B_\varepsilon(\tau(x))^c) f\left(\frac{p_{\tau(x),\varepsilon^c}(x)}{q(x)} \right) Q(dx).
\end{aligned}$$

Using the convexity of f and the definition of W, it can be checked that this is bounded below by

$$W f\left(\frac{1 - r_{\Pi,\varepsilon}}{W} \right) + (1 - W) f\left(\frac{r_{\Pi,\varepsilon}}{1 - W} \right).$$

Now inequality (1.4.5) can be inverted in some cases, as to provide a lower bound on $r_{\Pi,\varepsilon}$ and consequently also a lower bound on R_Π. A general technique, which is Corollary II.2 in Guntuboyina (2011), uses a first order approximation of the convex function $g(r) = Wf((1-r)/W)+(1-W)f(r/(1-W))$: for $0 \leqslant r_0 \leqslant 1-W$, we have $g'(r_0) \leqslant g'(1-W) = 0$, and hence

$$\int_\Theta D_f(\mathbb{P}_\theta \| Q)\Pi(d\theta) \geqslant g(r_{\Pi,\varepsilon}) \geqslant g(r_0) + g'(r_0)(r_{\Pi,\varepsilon} - r_0)$$

implies

$$r_{\Pi,\varepsilon} \geqslant r_0 + \frac{\int_\Theta D_f(\mathbb{P}_\theta \| Q)\Pi(d\theta) - g(r_0)}{g'(r_0)}. \tag{1.4.6}$$

Since Q was arbitrary, we also have

$$r_{\Pi,\varepsilon} \geqslant r_0 + \frac{\inf_{Q \ll \lambda} \int_\Theta D_f(\mathbb{P}_\theta \| Q)\Pi(d\theta) - g(r_0)}{g'(r_0)}. \tag{1.4.7}$$

To see how this may be used in practice, suppose that Π satisfies the condition $\Pi(B_\varepsilon(\theta_0)) \in \{0, 1/N\}$ for some $N \geqslant 1$ and every $\theta_0 \in \Theta$. In this case, $W = 1/N$. If $f(t) = t\log(t)$, so that D_f is the Kullback-Leibler divergence, and with $r_0 = (N-1)/(2N-1)$, we find $g'(r_0) = \log(1/N)$, $g(r_0) = \{(n-1)\log(n/(2n-1)) + n\log(n^2/(2n-1))\}/(2n-1)$ and

$$\begin{aligned} r_{\Pi,\varepsilon} &\geqslant 1 - \frac{\inf_{Q \ll \lambda} \int_\Theta D_f(\mathbb{P}_\theta \| Q)\Pi(d\theta) + \log(2N-1) - \log(N)}{\log(N)} \\ &\geqslant 1 - \frac{\inf_{Q \ll \lambda} \int_\Theta D_f(\mathbb{P}_\theta \| Q)\Pi(d\theta) + \log(2)}{\log(N)}. \end{aligned}$$

The existence of such a probability measure Π can be seen as depending on the metric entropy of Θ. Suppose for instance that the loss ℓ is a distance and let Θ_ε be a 2ε-net of Θ, which we assume to be finite. That is, for any $\theta, \theta' \in \Theta_\varepsilon$, either $\ell(\theta, \theta') > 2\varepsilon$ or $\theta = \theta'$. Then with $N = N_\varepsilon = |\Theta_\varepsilon|$ and $\Pi = \frac{1}{N}\sum_{\theta \in \Theta_\varepsilon} \delta_\theta$, we have $\Pi(B_\varepsilon(\theta_0)) \in \{0, 1/N\}$ for every $\theta_0 \in \Theta$ and the minimax and Bayes risks are bounded below by

$$\varepsilon\left(1 - \frac{\inf_{Q \ll \lambda} \int_\Theta D_f(\mathbb{P}_\theta \| Q)\Pi(d\theta) + \log(2)}{\log(N_\varepsilon)}\right).$$

In order to bound $\inf_{Q \ll \lambda} \int_\Theta D(\mathbb{P}_\theta \| Q) \Pi(d\theta)$, fix $\delta > 0$ and suppose there exists a finite set $\Theta'_\delta \subset \Theta$ such that for every $\theta \in \Theta$, $\exists \theta' \in \Theta'_\delta$ with $D(\mathbb{P}_\theta \| \mathbb{P}_{\theta'}) < \delta$. Then with $Q = \frac{1}{|\Theta'_\delta|} \sum_{\theta' \in \Theta'_\delta}$, we find

$$\inf_{Q \ll \lambda} \int_\Theta D(\mathbb{P}_\theta \| Q) \Pi(d\theta) \leqslant \sup_{\theta \in \Theta} D(\mathbb{P}_\theta \| Q)$$

and for every $\theta \in \Theta$,

$$\begin{aligned} D(\mathbb{P}_\theta \| Q) &= \mathbb{E}_{X \sim \mathbb{P}_\theta} \left[\log \left(\frac{p_\theta(X)}{\frac{1}{|\Theta'_\delta|} \sum_{\theta' \in \Theta'_\delta} p_{\theta'}(X)} \right) \right] \\ &\leqslant \log(|\Theta'_\delta| + \inf_{\theta' \in \Theta'_\delta} D(\mathbb{P}_\theta \| \mathbb{P}_{\theta'}) \\ &\leqslant \log(|\Theta'_\delta|) + \delta. \end{aligned}$$

We therefore obtained

$$R_\Pi \geqslant \varepsilon \left(1 - \frac{\log(|\Theta'_\delta|) + \delta + \log(2)}{\log(|\Theta_\varepsilon|)} \right). \tag{1.4.8}$$

The quantities $|\Theta'_\delta|$ and $|\Theta_\varepsilon|$, which are respectively covering and packing numbers for the Kullback-Leibler divergence and the ℓ distance, can be related to one another using inequalities between ℓ and the Kullback-Leibler divergence. With δ such that $\log(|\Theta'_\delta|) = \delta$ and ε satisfying $\log(|\Theta_\varepsilon|) \geqslant 4\delta + 2\log(2)$, we have then showed that $R_\Pi \geqslant \varepsilon/2$. See Yang and Barron (1999) for a more in-depth study and the analysis of particular cases.

CHAPTER II

NOTE ON REVERSE PINSKER INEQUALITIES

2.1 Abstract

A simple method is shown to provide optimal variational bounds on f-divergences with possible constraints on relative information extremums. Known results are refined or proved to be optimal as particular cases.

2.2 Introduction

This note is concerned with optimal upper bounds on relative entropy and other f-divergences in terms of the total variation distance and relative information extremums. When taking relative entropy as the f-divergence, such upper variational bounds have been referred to as *reverse Pinsker inequalities* (Sason and Verdú, 2016; Böcherer and Geiger, 2016). They are used in the optimal quantization of probability measures (Böcherer and Geiger, 2016) and have also appeared in Bayesian nonparametrics for controlling the prior probability of relative entropy neighbourhoods (see e.g. Lemma 8.2 of Ghosal et al. (2000)).

Our main theorem demonstrates a simple method that yields optimal "reverse Pinsker inequalities" for any f-divergence. This refines or shows the optimality of previously best known inequalities while avoiding arguments that are tuned to particular cases. In particular, Simic (2009a) uses a global upper bound on the Jensen function to bound relative entropy by a function of relative information extremums. Corollary 2.3.2 below refines their inequality to best possible. More recently, three different bounds on relative entropy involving the total variation distance have been proposed in Theorem 23 of Sason and Verdú (2016) in Theorem 7 of Verdú (2014) and in Theorem 1 of Sason (2015). Our results show that the inequalities of Sason and Verdú (2016) and Verdú (2014) are in fact optimal in related contexts. Another direct application of the method improves Theorem

34 in Sason and Verdú (2016), which is an upper bound on Rényi's divergence in terms of the variational distance and relative information maximum, while providing a simpler proof for this type of inequality. Vajda's well-known "range of values theorem" (see Vajda (1972); Liese and Vajda (2006); Vajda (2009); Kumar and Hunter (2004); Kumar and Chhina (2005)) is also recovered as an application.

The rest of the paper is organized as follows. Section 3.5.3 presents the definitions and main results. Examples with particular f-divergences are provided in section 2.4 and proofs are given in section 2.5.

2.3 Main results

Let (P, Q) be a pair of probability measures. It is assumed throughout that $P \ll Q$. Given a convex function $f : [0, \infty) \to (-\infty, \infty]$ such that $f(1) = 0$, the f-divergence between P and Q is defined as

$$D_f(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right]. \tag{2.3.1}$$

In particular, the relative entropy $D(P\|Q)$ and the total variation distance $D_{TV}(P, Q) = \sup_A |P(A) - Q(A)|$ correspond to the cases $f(t) = t\log(t)$ and $f(t) = \frac{1}{2}|t - 1|$ respectively.

For fixed $\delta, m \geqslant 0$ and $M \leqslant \infty$, we consider the set $\mathcal{A}(\delta, m, M)$ of all probability measure pairs (P, Q) respecting the conditions : $P \ll Q$,

$$\operatorname{ess\,inf} \frac{dP}{dQ} = m, \quad \operatorname{ess\,sup} \frac{dP}{dQ} = M \quad \text{and} \quad D_{TV}(P, Q) = \delta. \tag{2.3.2}$$

Here ess inf and ess sup represent the essential infimum and supremum taken with respect to Q.

The following theorem provides the best upper bound on the f-divergence over the class $\mathcal{A}(\delta, m, M)$ determined by (2.3.2).

Theorem 2.3.1. *If $\delta, m \geqslant 0$ and $M < \infty$ are such that $\mathcal{A}(\delta, m, M) \neq \emptyset$, then*

$$\sup_{(P,Q)\in\mathcal{A}(\delta,m,M)} D_f(P\|Q) = \delta\left(\frac{f(m)}{1-m} + \frac{f(M)}{M-1}\right). \tag{2.3.3}$$

Remark 2.3.1. If $m = 1$ or $M = 1$, then necessarily $\delta = 0$ and the right-hand side of (2.3.3) is to be interpreted as 0.

Remark 2.3.2. Theorem 2.3.1 generalizes Theorem 23 in Sason and Verdú (2016) with $f(t) = t\log(t)$ for the relative entropy: the upper bounds obtained are the same in this case. The concepts of the proofs also share similarities which are detailed in Remark 2.5.1 of Section 2.5.

We can obtain from Theorem 1 tight bounds for more general families of distributions. Consider for instance

$$\mathcal{B}(m, M) = \bigcup_{\delta\geqslant 0} \mathcal{A}(\delta, m, M) \tag{2.3.4}$$

and

$$\mathcal{C}(\delta) = \bigcup_{\substack{m\in[0,1]\\ M\in[1,\infty]}} \mathcal{A}(\delta, m, M). \tag{2.3.5}$$

Using the first family, Corollary 2.3.2 below provides the range of D_f as a function of relative information bounds.

Corollary 2.3.2. *If $m \geqslant 0$ and $M < \infty$ are such that $\mathcal{B}(m, M) \neq \emptyset$, then*

$$\sup_{(P,Q)\in\mathcal{B}(m,M)} D_f(P\|Q) = \frac{(M-1)f(m) + (1-m)f(M)}{M-m}. \tag{2.3.6}$$

Using the second family (2.3.5), we re-obtain Theorem 4 of Sason and Verdú (2016) (see also Lemma 11.1 in Basu et al. (2011)). Taking the union over possible values of δ also yields Vajda's well-known "range of values theorem" (see Liese and Vajda (2006); Vajda (1972, 2009); Kumar and Hunter (2004); Kumar and Chhina (2005)).

Corollary 2.3.3. *If* $0 \leqslant \delta \leqslant 1$, *then*

$$\sup_{(P,Q)\in\mathcal{C}(\delta)} D_f(P\|Q) = \delta\left(f(0) + \lim_{M\to\infty}\frac{f(M)}{M}\right). \tag{2.3.7}$$

2.4 Examples

This section lists applications to particular f-divergences and follows the standard definitions of Sason and Verdú (2016). The bounds obtained are compared to similar inequalities recently shown in the literature.

2.4.1 Relative entropy (Kullback-Leibler divergence)

The relative entropy corresponds to $f(t) = t\log(t)$ in (2.3.1) and is denoted $D(P\|Q)$. The results are more neatly stated in this case as functions of $a = \operatorname{ess\,inf}\frac{dQ}{dP} = M^{-1}$ and $b = \operatorname{ess\,sup}\frac{dQ}{dP} = m^{-1}$, assuming both quantities are well defined. Theorem 2.3.1 then shows

$$\sup_{(P,Q)\in\mathcal{A}(\delta,m,M)} D(P\|Q) = \delta\left(\frac{\log(a)}{a-1} + \frac{\log(b)}{1-b}\right).$$

In particular, the resulting upper bound on $D(P\|Q)$ is Theorem 23 of Sason and Verdú (2016). Letting $b \to \infty$ gives the related Theorem 7 of Verdú (2014) and the inequality presented therein is consequently optimal over $\bigcup_{0\leqslant m\leqslant 1}\mathcal{A}(\delta, m, M)$.

Also, Corollary 2.3.2 yields

$$\sup_{(P,Q)\in\mathcal{B}(m,M)} D(P\|Q) = \frac{(a-1)\log(b) + (1-b)\log(a)}{b-a}.$$

For comparison, Theorem I of Simic (2009a) (which also appears as Theorem I in Simic (2011) and is related to results in Simic (2009c,b)) provides the weaker upper bound

$$\frac{a\log(b) - b\log(a)}{b-a} + \log\left(\frac{b-a}{\log(b)-\log(a)}\right) - 1$$

on $D(P\|Q)$ over $(P,Q) \in \mathcal{B}(m,M)$ as an application of their "best possible global bound" for the Jensen functional.

2.4.2 Hellinger divergence of order α

Let $\alpha \in (0,1) \cup (1,\infty)$ and $f(t) = (t^\alpha - 1)/(\alpha - 1)$. The corresponding divergence is denoted $\mathcal{H}_\alpha(P\|Q)$. Theorem 1 shows in this case

$$\sup_{(P,Q)\in\mathcal{A}(\delta,m,M)} \mathcal{H}_\alpha(P\|Q) = \frac{\delta}{1-\alpha}\left(\frac{1-m^\alpha}{1-m} - \frac{M^\alpha - 1}{M-1}\right).$$

When $\alpha = 2$, $\mathcal{H}_\alpha = D_{\chi^2}$ is the χ^2 divergence and the above can be rewritten as

$$\sup_{(P,Q)\in\mathcal{A}(\delta,m,M)} D_{\chi^2}(P\|Q) = \delta(M-m).$$

For comparison, Example 6 of Theorem 5 in Sason and Verdú (2016) is the weaker inequality

$$D_{\chi^2}(P\|Q) \leqslant 2\delta \max\{M-1, 1-m\}.$$

2.4.3 Rényi's divergence

Also related is Rényi's α-divergence, defined as

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log(1 + (\alpha-1)\mathcal{H}_\alpha(P\|Q))$$

and which is a monotonous transform of $\mathcal{H}_\alpha$. Correspondingly we obtain

$$D_\alpha(P\|Q) \leqslant \frac{1}{\alpha-1}\log\left(1 + \delta\left(\frac{M^\alpha - 1}{M-1} - \frac{1-m^\alpha}{1-m}\right)\right).$$

Taking $m = 0$ recovers Theorem 34 of Sason and Verdú (2016). Their inequality, which is also appears in Theorem 3 of Sason and Verdú (2015) for $\alpha > 2$, is improved when $m > 0$.

2.5 Proofs

The starting point of our analysis is the following simple known application of convexity.

Lemma 2.5.1. *Let κ be a random variable with values in a bounded interval $I = [a, b]$, let $\varphi : I \to (-\infty, \infty]$ be a convex function and let $\overline{\alpha} = (b - \mathbb{E}[\kappa])/(b - a)$. Then*

$$\mathbb{E}[\varphi(\kappa)] \leqslant \overline{\alpha}\varphi(a) + (1 - \overline{\alpha})\varphi(b). \tag{2.5.1}$$

Proof. Let α be a non-negative random variable such that $\kappa = \alpha a + (1 - \alpha)b$. Then $\mathbb{E}[\alpha] = \overline{\alpha}$ and by convexity of φ we find

$$\mathbb{E}[\varphi(\kappa)] \leqslant \mathbb{E}[\alpha\varphi(a) + (1 - \alpha)\varphi(b)] = \overline{\alpha}\varphi(a) + (1 - \overline{\alpha})\varphi(b).$$

□

As a particular case, we obtain a bound on the total variation distance that is of use in the proof of Theorem 2.3.1.

Corollary 2.5.2. *If $m \geqslant 0$, $M < \infty$ and $(P, Q) \in \mathcal{B}(m, M)$, then*

$$D_{TV}(P, Q) \leqslant \frac{(M - 1)(1 - m)}{M - m}. \tag{2.5.2}$$

Proof. Lemma 2.5.1, applied with $\kappa = \frac{dP}{dQ}$, $\varphi(x) = |x - 1|$, $a = m$ and $b = M$, shows that

$$\begin{aligned} \frac{1}{2}\mathbb{E}_Q\left[\left|\frac{dP}{dQ} - 1\right|\right] &\leqslant \frac{1}{2}\left[\frac{M - 1}{M - m}|m - 1| + \frac{1 - m}{M - m}|M - 1|\right] \\ &= \frac{(M - 1)(1 - m)}{M - m}. \end{aligned}$$

□

We now proceed with the proof of Theorem 2.3.1.

Proof of Theorem 2.3.1. Let $(P, Q) \in \mathcal{A}(\delta, m, M)$. If $A = \left\{x \mid \frac{dP}{dQ}(x) \leqslant 1\right\}$, then $\delta = Q(A) - P(A)$ and we may write

$$D_f(P\|Q) = Q(A)\mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\middle|A\right] + Q(A^c)\mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\middle|A^c\right]. \tag{2.5.3}$$

To bound the first term on the right-hand side of (2.5.3), note that $\mathbb{E}_Q\left[\frac{dP}{dQ}\middle|A\right] = \frac{P(A)}{Q(A)}$ and that $x \in A$ implies $m \leqslant \frac{dP}{dQ}(x) \leqslant 1$. An application of Lemma 2.5.1, using the fact that $f(1) = 0$, therefore yields

$$\mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\middle|A\right] \leqslant \frac{1 - \frac{P(A)}{Q(A)}}{1 - m} f(m) = \frac{\delta f(m)}{Q(A)(1 - m)}. \tag{2.5.4}$$

The second term is similarly bounded as to obtain

$$\mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\middle|A^c\right] \leqslant \frac{\delta f(M)}{Q(A^c)(M - 1)}. \tag{2.5.5}$$

Together with (2.5.3), the inequalities (2.5.4) and (2.5.5) show that

$$D_f(P\|Q) \leqslant \delta\left(\frac{f(m)}{1 - m} + \frac{f(M)}{M - 1}\right) \tag{2.5.6}$$

whenever $(P, Q) \in \mathcal{A}(\delta, m, M)$.

We now show that the supremum of (2.3.3) indeed attains this bound. The cases $\delta = 0$ and $\delta = 1$ are easily treated as they correspond to equality or mutual singularity of P and Q. We can therefore assume $0 < \delta < 1$. Let $q = \frac{M-1}{M-m}$, $p = mq$, $t = \delta(M - m)[(M - 1)(1 - m)]^{-1}$ and consider the pair of discrete measures

$$\begin{cases} P_0 = (tp, t(1 - p), 1 - t), \\ Q_0 = (tq, t(1 - q), 1 - t). \end{cases} \tag{2.5.7}$$

Corollary 2.5.2 ensures $0 < t < 1$ and thus P_0 and Q_0 are probability measures. It is also straightforward to verify that $(P_0, Q_0) \in \mathcal{A}(\delta, m, M)$ with $t(q - p) = \delta$,

$p/q = m$ and $(1-p)/(1-q) = M$. Some algebraic manipulations then show

$$\begin{aligned} D_f(P_0, Q_0) &= tqf\left(\frac{p}{q}\right) + t(1-q)f\left(\frac{1-p}{1-q}\right) \\ &= \delta\left(\frac{f(m)}{1-m} + \frac{f(M)}{M-1}\right). \end{aligned}$$

□

Remark 2.5.1. A decomposition equivalent to (2.5.3) is also used in the proof of Theorem 23 in Sason and Verdú (2016) wherein $f(t) = t\log(t)$. They then proceed to obtain the upper bound (2.5.6) using the monotonicity of the function $t \mapsto t\log(t)/(1-t)$ (continuously extended at 0 and 1).

Proof of Corollary 2.3.2. Combining Corollary 2.5.2 with equation (2.3.3) of Theorem 2.3.1 yields the upper bound. To see that the supremum attains this bound, let $\delta \to (M-1)(1-m)/(M-m)$ in (2.3.3). □

Proof of Corollary 2.3.3. Some care has to be taken when considering the elements of $\mathcal{A}(\delta, 0, \infty)$. To see that the right-hand side of (2.3.7) also upper bounds the elements of this set, we again use the decomposition (2.5.3). The first term is treated as in (2.5.4). For the second term, let $\frac{dP}{dQ} \wedge K = \min\{\frac{dP}{dQ}, K\}$. By Fatou's lemma and Lemma 2.5.1, using that $f(1) = 0$,

$$\begin{aligned} \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\middle|A^c\right] &\leqslant \liminf_{K\to\infty} \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\wedge K\right)\middle|A^c\right] \\ &\leqslant \liminf_{K\to\infty} \frac{\mathbb{E}_Q\left[\frac{dP}{dQ}\wedge K\middle|A^c\right]-1}{K-1} f(K). \end{aligned}$$

By the monotone convergence theorem,

$$\lim_{K\to\infty} \mathbb{E}_Q\left[\frac{dP}{dQ}\wedge K\middle|A^c\right] = \frac{P(A^c)}{Q(A^c)}$$

and hence

$$\mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\middle|A^c\right] \leqslant \frac{\delta}{Q(A^c)} \lim_{M\to\infty} \frac{f(M)}{M-1}.$$

We note that $\lim_{M\to\infty} \frac{f(M)}{M-1}$ exists by convexity of f and can be infinite. The required upper bound on $D_f(P\|Q)$ is then obtained as in the proof of Theorem 2.3.1.

To see that the upper bound is attained, it suffices to let $M \to \infty$ in Theorem 2.3.1. $\square$

CHAPTER III

BAYESIAN NONPARAMETRICS FOR DIRECTIONAL STATISTICS

3.1 Abstract

We introduce a density basis of the trigonometric polynomials that is suitable to mixture modelling. Statistical and geometric properties are derived, suggesting it as a circular analogue to the Bernstein polynomial densities. Nonparametric priors are constructed using this basis and a simulation study shows that the use of the resulting Bayes estimator may provide gains over comparable circular density estimators previously suggested in the literature.

From a theoretical point of view, we propose a general prior specification framework for density estimation on compact metric space using sieve priors. This is tailored to density bases such as the one considered herein and may also be used to exploit their particular shape-preserving properties. Furthermore, strong posterior consistency is shown to hold under notably weak regularity assumptions and adaptative convergence rates are obtained in terms of the approximation properties of positive linear operators generating our models.

3.2 Introduction

There is increasing interest in the statistical analysis of non-euclidean data, such as data lying on a circle, on a sphere or on a more complex manifold or metric space. Applications range from the analysis of seasonal and angular measurements to the statistics of shapes and configurations (Jammalamadaka and SenGupta, 2001; Bhattacharya and Bhattacharya, 2012). In bioinformatics, for instance, an important problem is that of using the chemical composition of a protein to predict the conformational angles of its backbone (Al-Lazikani et al., 2001). Bayesian nonparametric methods, accounting for the wrapping of angular data, have been successfully applied in this context (Lennox et al., 2009, 2010).

Directional statistics deals in particular with univariate angular data and provides basic building blocks for more complex models. Among the most commonly used model for the probability density function of a circular random variable is the von Mises density defined by

$$u \mapsto \exp(\kappa \cos(u - \mu))/(2\pi I_0(\kappa)),$$

where μ is the circular mean, $\kappa > 0$ is a shape parameter and I_0 is the modified Bessel function of the first kind and order 0. This function is nonnegative, 2π-periodic and integrates to one on the interval $[0, 2\pi)$. It can be regarded a circular analogue to normal distribution (Jammalamadaka and SenGupta, 2001) (see also Coeurjolly and Le Bihan (2012) for a comparison with the geodesic normal distribution). Mixtures of von Mises densities and other log-trigonometric densities are also frequently used (Kent, 1983). Another natural approach is to model circular densities using trigonometric polynomials

$$u \mapsto \frac{1}{2\pi} + \sum_{k=1}^{n} (a_k \cos(ku) + b_k \sin(ku)). \tag{3.2.1}$$

These densities have tractable normalizing constants, but the coefficients a_k and b_k must be constrained as to ensure nonnegativity (Fejér, 1916; Fernández-Durán, 2004).

For a review of common circular distributions, see Mardia and Jupp (2000); Jammalamadaka and SenGupta (2001). Notable Bayesian approaches to directional statistics problems include Ghosh and Ramamoorthi (2003b); McVinish and Mengersen (2008); Ravindran and Ghosh (2011); Hernandez-Stumpfhauser et al. (2017).

In this paper, we introduce a basis of the trigonometric polynomials (3.2.1) consisting only of probability density functions. Properties shown in Section 3.3, such as its shape-preserving properties, suggest it as a circular analogue to the

Bernstein polynomial densities and we argue that it is particularly well suited to mixture modelling. In Section 3.4, we use this basis to devise nonparametric priors on the space of bounded circular densities. We compare their posterior mean estimates to other density estimation methods based on the usual trigonometric representation (3.2.1) in Section 3.5.

An important aspect of nonparametric prior specification is the posterior consistency property, which entails almost sure convergence (in an appropriate topology) of the posterior mean estimate. In Section 3.4.2, we thus develop a general prior specification framework that immediately provides consistency of a class of sieve priors for density estimation on compact metric spaces. Particular instances of this framework appeared previously in the literature. For instance, Petrone and Wasserman (2002) obtained consistency of the Bernstein-Dirichlet prior on the set of continuous densities on the interval $[0, 1]$. More recently Xing and Ranneby (2009) (see also Walker (2004); Lijoi et al. (2005)) have obtained a simple condition for models of this kind ensuring consistency on the Kullback-Leibler support of the prior. As an application, they quickly revisit the problem of Petrone and Wasserman (2002) but without discussing what contains the Kullback-Leibler support. Our main contribution here is the proof that the Kullback-Leibler support of the priors specified in our framework contains every bounded density. Furthermore, we show in Section 3.4.4 how our framework may be used to obtain posterior contraction rates. The results are related to those of Ghosal (2001); Kruijer and van der Vaart (2008) in the case of the Bernstein-Dirichlet prior but are stated with more generality. They express posterior contraction rates in terms of a balance between the dimension of the sieves and their approximation properties, as they are accounted for by a sequence of positive linear approximation operators.

3.3 *De la Vallée Poussin* mixtures for circular densities

3.3.1 The basis

We propose the basis $\mathcal{B}_n$ for 2π-periodic densities of circular random variables given by

$$C_{j,n}(u) = \frac{2^{2n}}{2\pi\binom{2n}{n}} \left(\frac{1 + \cos\left(u - \frac{2\pi j}{2n+1}\right)}{2} \right)^n, \quad u \in \mathbb{R}, \quad j = 0, \ldots, 2n, \qquad (3.3.1)$$

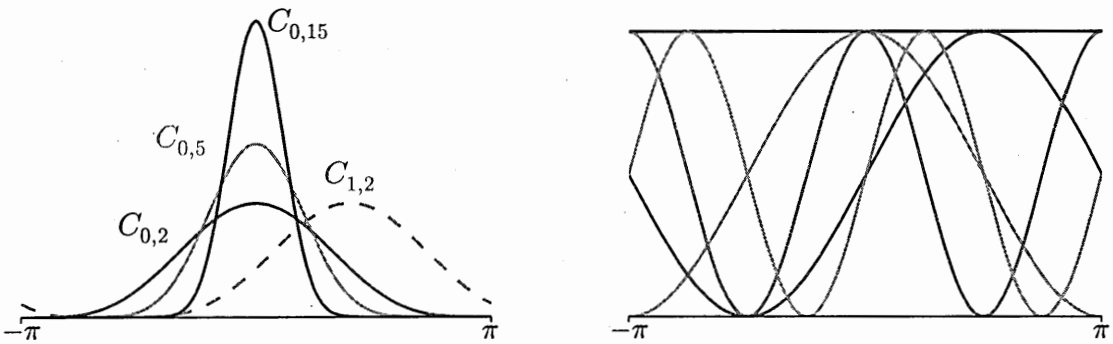


Figure 3.1: Comparison between De la Vallée Poussin basis densities (left) and the usual trigonometric basis 1, $\cos(x)$, $\sin(x)$, ... (right).

The rescalings $C^*_{j,n} = (2\pi/(2n+1))C_{j,n}$, $j = 0, \ldots, 2n$, were considered in Róth et al. (2009) in the context of *Computer Aided Geometric Design (CAGD)*. It was shown therein to actually form a basis for the vector space of trigonometric polynomials (of order at most $n \geqslant 1$) given by

$$\mathcal{V}_n = \operatorname{span}\{1, \cos u, \sin u, \ldots, \cos nu, \sin nu\}.$$

One important property of these rescalings to the CAGD community is that the resulting basis forms a partition of unity, meaning that $\sum_{j=0}^{2n} C^*_{j,n}(u) = 1$, for all $u \in \mathbb{R}$. The function $\omega_n = 2\pi C_{0,n}$ is the so-called *De la Vallée Poussin* kernel which has been studied by Pólya and Schoenberg (1958) and $C_{0,n}$ has also been refered to as Cartwright's power of cosine distribution Cartwright (1963).

We argue here that $\mathcal{B}_n$ provides an interesting model for densities of circular random variables, representing an angle or located on the circumference of a circle. Here is a formal definition of the *angular domain* on which we work.

Circular random variables take their values on a circle $\mathbb{S}^1$, which we identify to the real line modulo 2π. We therefore write $\mathbb{S}^1 = \mathbb{R} \pmod{2\pi}$, so that $\mathbb{S}^1$ consists of equivalence classes $\{x + 2\pi k : k \in \mathbb{Z}\}$ and is represented by any half-open interval of length 2π. In the following, we do not distinguish equivalence classes from their representatives. We endow $\mathbb{S}^1$ with the *angular distance* d defined as $d_{\mathbb{S}^1}(u, v) = \min_{k\in\mathbb{Z}} |u - v + 2\pi k|$. By the embedding $\theta \mapsto e^{i\theta}$ of $\mathbb{S}^1$ as the unit circle of the complex plane $\mathbb{C}$, the angular distance $d_{\mathbb{S}^1}$ becomes the arc length distance. For instance, an interval $[a, b) \subset \mathbb{S}^1$, $b - a < 2\pi$, can be viewed as an arc of length $b - a$ on the unit circle.

The following result gives elementary properties of the distributions corresponding to the densities in $\mathcal{B}_n$.

Theorem 3.3.1. *The random variables on* $\mathbb{S}^1$ *given by* $U_j = U + \frac{2\pi j}{2n+1}$, $j = 0, \ldots, 2n$, *where* $U = (1 - 2V)\cos^{-1}(1 - 2W)$, *with* V *and* W *independently distributed,* $V \sim Ber(1/2)$ *and* $W \sim Beta(1/2, 1/2 + n)$, *have* (3.3.1) *as densities. Furthermore, by letting* $Z_j = e^{iU_j}$ *be the corresponding random variable on the unit circle of* $\mathbb{C}$, *we have*

$$\mathbb{E}(Z_j^p) = \begin{cases} \frac{\binom{2n}{n-p}}{\binom{2n}{n}} e^{i\frac{2\pi jp}{2n+1}}, & \text{if } p \in \{-n, \ldots, n\}, \\ 0 & \text{if } p \in \mathbb{Z} \setminus \{-n, \ldots, n\}. \end{cases} \tag{3.3.2}$$

Proof. The first part is a straightforward application of the change of variables formula. For the integer moments, we have the equality $\mathbb{E}(Z_j^p) = e^{i\frac{2\pi jp}{2n+1}}\mathbb{E}(Z_0^p)$. Using the identity

$$C_{0,n}(u) = \frac{2^{2n}}{2\pi\binom{2n}{n}} \cos^{2n}(u/2), \quad u \in [0, 2\pi), \tag{3.3.3}$$

and letting $S \sim \mathcal{U}(\mathbb{S}^1)$, we find

$$\mathbb{E}\left(Z_0^p\right) = \frac{1}{\binom{2n}{n}} \sum_{k=0}^{2n} \binom{2n}{k} \mathbb{E}(e^{-i(n-k-p)S}) = \begin{cases} \frac{\binom{2n}{n-p}}{\binom{2n}{n}}, & \text{if } p \in \{-n, \ldots, n\}, \\ 0 & \text{if } p \in \mathbb{Z} \setminus \{-n, \ldots, n\}. \end{cases}$$

□

The above integer moments (3.3.2) are also known as the Fourier coefficients in Feller (1971, p. 631) and as trigonometric moments in the directional statistics jargon, see for instance Mardia and Jupp (2000), Jammalamadaka and SenGupta (2001) and recently Coeurjolly and Le Bihan (2012). From the result for $p = 1$, we get that the mean direction of the j^{th} component is $e^{i\frac{2\pi jp}{2n+1}}$ with the so-called *circular variance* equal to $1/(n+1)$.

3.3.2 The circular density model

Let Δ_{2n} be the $2n$-dimensional simplex $\Delta_{2n} = \{(c_0, \ldots, c_{2n}) \in [0,1]^{2n+1} : c_0 + \cdots + c_{2n} = 1\}$. Our model consists in mixtures of the form

$$C_n(u; c_0, \ldots, c_{2n}) = \sum_{j=0}^{2n} c_j C_{j,n}(u), \quad u \in \mathbb{R}, \tag{3.3.4}$$

with $(c_0, \ldots, c_{2n}) \in \Delta_{2n}$, and $n \geqslant 0$. Let $\mathcal{C}_n$, $n \geqslant 0$, represent the set of mixtures obtained this way; our model is therefore

$$\mathcal{C} = \bigcup_{n \geqslant 0} \mathcal{C}_n. \tag{3.3.5}$$

We now give a characterization of the model in terms of trigonometric polynomials. We use the following *degree elevation* lemma, which is a reformulation of Róth et al. (2009, Theorem 6).

Lemma 3.3.2 (Degree elevation formula). *Each $C_{j,n} \in \mathcal{B}_n$ given by* (3.3.1) *can be expressed as*

$$C_{j,n}(u) = \sum_{\ell=0}^{2(n+r)} d_{j,\ell}^{n,r} C_{\ell,n+r}(u), \tag{3.3.6}$$

with

$$d_{j,\ell}^{n,r} = \frac{1}{2(n+r)+1} \left\{ 1 + \frac{2\binom{2(n+r)}{n+r}}{\binom{2n}{n}} \sum_{k=0}^{n-1} \frac{\binom{2n}{k}}{\binom{2(n+r)}{k+r}} \cos\left(\tfrac{2(n-k)\pi\ell}{2(n+r)+1} - \tfrac{2(n-k)\pi j}{2n+1} \right) \right\}, \tag{3.3.7}$$

for $\ell \in \{0, 1, \ldots, 2(n+r)\}$, and $r \geqslant 0$.

To give the characterization, let $\mathcal{D}_n \subset \mathcal{V}_n$ be the subset of trigonometric polynomial densities (of order at most $n \geqslant 1$), and let $\mathcal{D}_n^+ \subset \mathcal{D}_n$ be the positive ones.

Theorem 3.3.3 (Characterization). *We have $\mathcal{C} = \bigcup_{n \geqslant 0} \{\mathcal{B}_n \cup \mathcal{D}_n^+\}$.*

Proof. If $C_n \in \mathcal{C}_n \cap \mathcal{B}_n^c$, then we have $C_n(u) > 0$ for all u, and this shows $\mathcal{C} \subset \bigcup_{n \geqslant 0} \{\mathcal{B}_n \cup \mathcal{D}_n^+\}$. For the converse inclusion, let $C_n \in \mathcal{D}_n^+$, be a positive trigonometric polynomial density, that is, $C_n(u) = \sum_{j=0}^{2n} c_j^n C_{j,n}(u) > 0$, for all $u \in \mathbb{S}^1$, with $\sum_{j=0}^{2n} c_j^n = 1$. Some of the c_j^n's may be negative here. However, by the degree elevation lemma we have

$$C_n(u) = \sum_{\ell=0}^{2(n+r)} \left\{ \sum_{j=0}^{2n} c_j^n d_{j,\ell}^{n,r} \right\} C_{\ell,n+r}(u),$$

with $d_{j,\ell}^{n,r}$ given by (3.3.7). The resulting coefficients $c_\ell^{n+r} = \sum_{j=0}^{2n} c_j^n d_{j,\ell}^{n,r}$ also have the property $\sum_{\ell=0}^{2(n+r)} c_\ell^{n+r} = 1$, and so it remains to show that there is some $r \geqslant 0$ such that $c_\ell^{n+r} \geqslant 0$, for every $\ell = 0, \ldots, 2(n+r)$. To see this, use (3.3.3) and the binomial identity to write

$$C_n\left(\tfrac{2\pi\ell}{2(n+r)+1}\right) = \frac{1}{2\pi} \left\{ 1 + \frac{2}{\binom{2n}{n}} \sum_{k=0}^{n-1} \binom{2n}{k} \sum_{j=0}^{2n} c_j^n \cos\left(\tfrac{2(n-k)\pi\ell}{2(n+r)+1} - \tfrac{2(n-k)\pi j}{2n+1} \right) \right\}.$$

After some manipulations, and using the fact that $k \mapsto \binom{2(n+r)}{k+r}$ is increasing on $\{0, \ldots, n-1\}$, we find

$$\begin{aligned}\left|\tfrac{2(n+r)+1}{2\pi} c_\ell^{n+r} - C_n\left(\tfrac{2\pi\ell}{2(n+r)+1}\right)\right| &\leqslant \alpha_1(n)\left(\sum_{k=0}^{n-1}\binom{2n}{k}\left|\frac{\binom{2(n+r)}{n+r}}{\binom{2(n+r)}{k+r}} - 1\right|\right) \\ &\leqslant \alpha_2(n)\left(\frac{\binom{2(n+r)}{n+r}}{\binom{2(n+r)}{r}} - 1\right),\end{aligned}$$

where $\alpha_1(n), \alpha_2(n) > 0$. A final calculation shows that

$$\frac{\binom{2(n+r)}{n+r}}{\binom{2(n+r)}{r}} - 1 = \frac{(2n+r)(2n+r-1)\cdots(n+r+1)}{(n+r)(n+r-1)\cdots(r+1)} - 1 \leqslant (1+n/r)^n - 1.$$

Since $C_n \in \mathcal{D}_n^+$ is positive by assumption, this shows that for large enough r, we have $c_\ell^{n+r} > 0$, for every $\ell = 0, \ldots, 2(n+r)$, and therefore $C_n \in \mathcal{C}$. □

As mentioned in the introduction, a criticism made by Ferreira et al. (2008) concerning the nonnegative trigonometric polynomials proposed by Fernández-Durán (2004) and Fernández-Durán (2007) is that "approximating a function (using nonnegative trigonometric polynomials) often results in a wiggly approximation, unlikely to be useful in most real applications".

In the following, we define the notion of *cyclic variations* to formalize "wiggliness" and show that it can be controlled using our basis.

One way of quantifying "wiggliness" was discussed by Pólya and Schoenberg (1958) via the cyclic variations. For a finite sequence $x = (x_1, \ldots, x_m)$, $m \geqslant 2$, denote by $v(x)$ the number of sign changes (from positive to negative or vice versa) in the terms of the sequence. Denote by $\mathring{v}(x) = v(x_i, x_{i+1}, \ldots, x_m, x_1, x_2, \ldots, x_{i-1}, x_i)$, $x_i \neq 0$, the *cyclic variation* of the sequence, with $\mathring{v}(x) = 0$ if $x = 0$. This is well defined because $\mathring{v}$ does not depend on the particular index i such that $x_i \neq 0$. Notice that the value of $\mathring{v}$ is always an even number not exceeding m. The sequence x is said to be *periodically unimodal* if $\mathring{v}(\mathring{\Delta}x) = 2$, where

$\mathring{\Delta}x = (x_2 - x_1, \ldots, x_m - x_{m-1}, x_1 - x_m)$. For a function $f : \mathbb{S}^1 \to \mathbb{R}$, we make use of the notation

$$\mathring{v}(f) = \sup\{\mathring{v}(f(x_i)_{i=1}^m) : 0 \leqslant x_1 < x_2 < \cdots < x_m < 2\pi,\, m \geqslant 2\},$$

and $Z(f) = \#\{x \in [0, 2\pi) : f(x) = 0\}$. Similarly to the discrete case, such a function f is said to be *periodically unimodal*, also called *periodically monotone* by Pólya and Schoenberg (1958), if $\mathring{v}(f') = 2$, provided f' exists (a more general definition without the differentiability assumption is given in the latter paper but is not needed in our case).

We have the following results.

Theorem 3.3.4. *For $C_n = \sum_{j=0}^{2n} c_j C_{j,n} \in \mathcal{C}_n$, let $c = (c_0, \ldots, c_{2n}) \in \Delta_{2n}$. We have*

(i)

$$\mathring{v}(C_n - \alpha) \leqslant Z(C_n - \alpha) \leqslant \mathring{v}\left(\frac{2n+1}{2\pi}c - \alpha\right), \quad \text{for all } \alpha \geqslant 0.$$

(ii) A bound for the total variation of C_n is given by

$$\mathrm{TV}(C_n) := \int_0^{2\pi} |C_n'(u)|\, du \leqslant \frac{2n+1}{2\pi} \sum_{j=0}^{2n} |c_{j+1} - c_j| \leqslant (2n+1)/\pi,$$

where $c_{2n+1} = c_0$.

(iii) If $c = (c_0, \ldots, c_{2n})$ is periodically unimodal, then C_n is also periodically unimodal.

Proof. The proof of *(i)* follows by Pólya and Schoenberg (1958, Lemma 3) by noticing that

$$C_n(u) - \alpha = \sum_{j=0}^{2n} \left\{\frac{c_j}{2\pi} - \frac{\alpha}{2n+1}\right\} \omega_n\left(u - \frac{2\pi j}{2n+1}\right), \quad u \in \mathbb{S}^1,$$

with $\omega_n = 2\pi\, C_{0,n}$ the *De la Vallée Poussin* kernel. Their result says (in this case) that $Z(C_n - \alpha) \leqslant \mathring{v}\,(c_j/2\pi - \alpha/(2n+1))_{j=0}^{2n}$, which implies *(i)*.

To show *(ii)*, let $P_n : \mathbb{S}^1 \to \mathbb{R}$ be the continuous and 2π-periodic, piecewise linear interpolation of the points $(2\pi j/(2n+1), (2n+1)c_j/2\pi) \in \mathbb{S}^1 \times \mathbb{R}$, $j \in \{0, \ldots, 2n\}$. For definiteness,

$$P_n(u) = \sum_{j=0}^{2n} c_j L_j(u), \quad u \in \mathbb{S}^1, \tag{3.3.8}$$

where $L_j(u) = 0 \vee \frac{2n+1}{2\pi}(1 - \frac{2n+1}{2\pi} d_{\mathbb{S}^1}(u, \frac{2\pi j}{2n+1}))$. By *(i)* and the Banach Indicatrix Theorem, see Benedetto and Czaja (2009), we have

$$\begin{aligned} \mathrm{TV}(C_n) = \int_0^\infty Z(C_n - \alpha)\, d\alpha &\leqslant \int_0^\infty \mathring{v}\left(\frac{2n+1}{2\pi} c - \alpha\right) d\alpha, \\ &\leqslant \int_0^\infty Z(P_n - \alpha)\, d\alpha \\ &= \mathrm{TV}(P_n) = \frac{2n+1}{2\pi} \sum_{j=0}^{2n} |c_{j+1} - c_j|. \end{aligned}$$

Now a (sharp) bound is easily found for the last sum by $\sum_{j=0}^{2n} |c_{j+1} - c_j| = \|(c_1, \ldots, c_{2n+1}) - (c_0, \ldots, c_{2n})\|_1 \leqslant 2$, which leads to the assertion $\mathrm{TV}(C_n) \leqslant (2n+1)/\pi$.

For *(iii)*, we assume $\mathring{v}(\mathring{\Delta} c) = 2$ and we want to show that $\mathring{v}(C_n') = 2$. First, if $\mathring{v}(C_n') = 0$ then C_n' is either nonnegative or nonpositive. By continuity of C_n', we have $0 = C_n(2\pi) - C_n(0) = \int_0^{2\pi} C_n'(u)\, du$, which implies $C_n'(u) = 0$, for all $u \in [0, 2\pi)$, and this gives $c_i = 1/(2n+1)$, $i = 0, \ldots, 2n$. Thus, $\mathring{v}(C_n') = 2k$, for some $1 \leqslant k \leqslant n$. The unit circle $\mathbb{S}^1$ can therefore be partitioned into $2k$ open arcs $A_1, \ldots, A_{2k}$ with $(-1)^j C_n$ being nondecreasing on A_j, $j = 1, \ldots, 2k$ and with (anticlockwise) end points $a_1, \ldots, a_{2k}$ (listed in anticlockwise order) being interlaced local minima $\{a_1, a_3 \ldots, a_{2k-1}\}$ and maxima $\{a_2, \ldots, a_{2k}\}$ of C_n. Assume $k > 1$ and without loss of generality $a_2 \leqslant a_4$. Let $m = \max\{a_1, a_3\}$. By the monotonicity of C_n on each arc, each of which being a connected set (relatively to

the topology induced by the angular distance d), the Intermediate Value Theorem gives $Z(C_n - \alpha) > 2$ for all $\alpha \in (m, a_2)$. By the same argument, using the fact that $\mathring{v}(\mathring{\Delta}c) = 2$, we obtain

$$\mathring{v}\left(\frac{2n+1}{2\pi}c - \alpha\right) = \begin{cases} 2, & \text{if } \alpha \in (\min(c), \max(c)), \\ 0 & \text{otherwise,} \end{cases}$$

contradicting *(i)*, and this implies $k = 1$. □

3.4 Prior specification

3.4.1 Circular density prior

Our prior Π on the space $\mathbb{F} = \mathbb{F}(\mathbb{S}^1)$ of bounded circular densities, parametrized by a Dirichlet process $\mathcal{D}$ and a distribution ρ on $\{1, 2, 3, \dots\}$, is induced by the random density

$$\sum_{j=0}^{2N} \mathcal{D}(R_{j,N}) C_{j,N}, \quad N \sim \rho, \tag{3.4.1}$$

where $R_{j,n} = \left[\frac{\pi(2j-1)}{2n+1}, \frac{\pi(2j+1)}{2n+1}\right) \subset \mathbb{S}^1$. If $\mathcal{D}$ has a base probability measure G and a concentration parameter $M > 0$, then

$$\Pi(B) = \sum_{n \geqslant 0} \rho(n) \Pi_n(B \cap \mathcal{C}_n), \quad B \in \mathcal{B}, \tag{3.4.2}$$

where $\Pi_n = \Pi_{\Delta_{2n}} \circ l_n^{-1}$, $\Pi_{\Delta_{2n}}$ is the Dirichlet distribution of parameters $MG(R_{j,n})$, $j = 0, 1, \dots 2n$, and where $l_n : \Delta_{2n} \ni (c_0, \dots, c_{2n}) \mapsto \sum_{j=0}^{2n} c_j C_{j,n} \in \mathcal{C}_n$.

Strong posterior consistency is obtained using Theorem 3.4.3 of Section 3.4.2. The theorem requires the conditional distributions Π_n to have full support on $\mathcal{C}_n$, that $0 < \rho(n) < ce^{-Cn}$ for some $c, C > 0$, and that proper approximation properties of the sieves $\mathcal{C}_n$ are assessed by a sequence $T_n : L^1(\mathbb{M}) \to L^1(\mathbb{M})$ of linear operators, mapping densities to densities, such that $T_n(\mathbb{F}) = \mathcal{C}_n \subset \mathbb{F}$. Here we let T_n be

defined by

$$T_n f = \sum_{j=0}^{2n} \int_{R_{j,n}} f(u)du\, C_{j,n}. \tag{3.4.3}$$

The only condition of the theorem that is not readily verified is given in the following lemma.

Lemma 3.4.1. *For every continuous function f on $\mathbb{S}^1$, $\|T_n f - f\|_\infty \to 0$.*

Proof. We use Lemma 3.9.1, in the appendix (a result is similar to that of Lorentz (1986, Theorem 1.2.1)), which gives three sufficient conditions *(i)* – *(iii)* for uniform convergence. We denote $d_{\mathbb{S}^1}(u, R_{j,n}) = \inf_{v \in R_{j,n}} d(u, v)$, and $\mathrm{diam}(R_{j,n}) = \sup_{u,v \in R_{j,n}} d_{\mathbb{S}^1}(u, v)$. Here *(i)* is immediate by $\mathrm{diam}(R_{j,n}) = 2\pi/(2n+1)$, $j = 0, \ldots, 2n$, and *(iii)* follows from the partition of unity property of $\frac{2\pi}{2n+1} C_{j,n}$. Assumption *(ii)* follows since $C_{0,n}$ is unimodal with mode at 0, and $d_{\mathbb{S}^1}(u, R_{j,n}) \geqslant \delta > 0$ implies

$$C_{j,n}(u) = C_{0,n}\left(d_{\mathbb{S}^1}\left(u, \frac{2\pi j}{2n+1}\right)\right) \leqslant C_{0,n}\left(d_{\mathbb{S}^1}\left(u, R_{j,n}\right)\right) \leqslant C_{0,n}(\delta),$$

therefore $\sum_{j: d_{\mathbb{S}^1}(u, R_{j,n}) \geqslant \delta} \frac{2\pi}{2n+1} C_{j,n}(u) \leqslant 2\pi C_{0,n}(\delta) \to 0$, $n \to \infty$, uniformly over $u \in \mathbb{S}^1$. □

The prior may be interpreted similarly as the Bernstein-Dirichlet prior of Petrone (1999). Conditionally on a fixed n, the random histogram $H_n = \frac{2n+1}{2\pi} \sum_{j=0}^{2n} c_{j,n} \mathbb{1}_{R_{j,n}}$ is immediately understood through the Dirichlet distribution on $(c_{0,n}, \ldots, c_{2n,n})$. Since $\sum_{j=0}^{2n} c_{j,n} C_{j,n} = T_n H_n$, the following proposition together with Lemma 3.4.1 shows that the finite mixture (3.4.1) may be seen as a smooth, variation diminishing approximation to H_n.

Proposition 3.4.2 (Variation diminishing property)**.** *For every density f on $\mathbb{S}^1$, continuous on $R_{j,n}$, $j = 0, \ldots, 2n$, we have $\mathring{v}(T_n f - \alpha) \leqslant \mathring{v}(f - \alpha)$ for all $\alpha > 0$.*

Proof. This is a straightforward consequence of Theorem 3.3.4 *(i)*. Indeed, by continuity of f, the Mean Value Theorem says that $P_f(R_{j,n}) = \frac{2\pi}{2n+1} f(u_j)$, for some $u_j \in R_{j,n}$, $j = 0, \ldots, 2n$. It follows that

$$\mathring{v}(T_n f - \alpha) \leqslant \mathring{v}\left((P_f(R_{0,n}), \ldots, P_f(R_{2n,n})) - \alpha\right) \leqslant \mathring{v}\left(f - \alpha\right), \quad \alpha > 0.$$

□

3.4.2 Strong posterior consistency

We show the strong posterior consistency of a general class of priors for bounded density spaces on compact metric spaces. These include sieve priors such as (3.4.2), as well as a class of Dirichlet process location mixtures (see §3.4.3). In contrast with Bhattacharya and Dunson (2012), who also obtained general strong consistency result, we consider a prior specification framework, with a different applicability, that does not require continuity and positivity assumptions on the true density from which observations are made.

Here, strong consistency on $\mathbb{F}$ means that if $X_1, \ldots, X_n$ are independent random variables and identically distributed according to the probability distribution P_{f_0} with density $f_0 \in \mathbb{F}$, denoted $(X_i)_{i \geqslant 1} \sim P_{f_0}^{(\infty)}$, then for all $\varepsilon > 0$,

$$\Pi\left(\left\{f \in \mathbb{F} : \int |f - f_0| < \varepsilon\right\} \mid (X_i)_{i=1}^n\right) \to 1, \quad P_{f_0}^{(\infty)}\text{-a.s.} \tag{3.4.4}$$

The general framework is the following. Suppose $\mathbb{F}$ is the space of all bounded densities with respect to some finite measure μ on a compact metric space $(\mathbb{M}, d)$. Let $T_n : L^1(\mathbb{M}) \to L^1(\mathbb{M})$, $n \in \mathbb{N}$, be a sequence of linear operators mapping densities to densities. Consider a model having the form $\mathcal{C} = \cup_{n \geqslant 0} \mathcal{C}_n$, with $\mathcal{C}_n := T_n(\mathbb{F}) \subset \mathbb{F}$. Let $\mathfrak{B}$ be the Borel σ-algebra of $\mathbb{F}$ for the L^1 metric and let $\mathfrak{B}_n$ be the restriction of $\mathfrak{B}$ to $\mathcal{C}_n$, $n \geqslant 0$. A prior Π on $\mathbb{F}$ can be specified through priors

Π_n on $(\mathcal{C}_n, \mathfrak{B}_n)$ and a distribution ρ on $n \in \{0, 1, 2, \dots\}$ as

$$\Pi(B) = \sum_{n \geqslant 0} \rho(n) \Pi_n(B \cap \mathcal{C}_n), \quad B \in \mathfrak{B}. \tag{3.4.5}$$

In Theorem 3.4.3 below, we give simple conditions on Π_n, T_n and ρ, in this framework, ensuring strong posterior consistency on all of $\mathbb{F}$. The proof is given in the appendix.

Theorem 3.4.3. *Let $\mathbb{F}$, Π_n, Π and T_n be as above. Suppose that $T_n(\mathbb{F}) \subset \mathbb{F}$ are of finite dimensions bounded by an increasing sequence $d_n \in \mathbb{N}$, and also that $\|T_n f - f\|_\infty \to 0$, $n \to \infty$, for every continuous function f on $\mathbb{M}$. If $0 < \rho(n) < ce^{-Cd_n}$, for some $c > 0$, $C > 0$ and if Π_n has support $T_n(\mathbb{F})$, then the posterior distribution of Π is strongly consistent on $\mathbb{F}$.*

The proof is in Appendix 3.8.1.

Remark 3.4.1. The result still holds when the space $\mathbb{F}$ is constrained such as being some convex subset of bounded densities containing at least one density that is bounded away from zero or a star-shaped subset around such a density (e.g. $\mathbb{F}$ may be a set of bounded unimodal densities or a set of continuous multivariate copula densities). The precise conditions required on $\mathbb{F}$ are stated at the beginning of Appendix 3.7.1.

3.4.3 Relationship with Dirichlet Process Mixtures

Here we consider Dirichlet Process location Mixtures on $\mathbb{F}$ induced by the random density

$$f = \int_{\mathbb{M}} f(\cdot \mid \mu, n) \mathcal{D}(d\mu), \tag{3.4.6}$$

where $\{f(\cdot \mid \mu, n) \mid \mu \in \mathbb{M}\} \subset \mathbb{F}$ are families of densities, $\mathcal{D}$ is a Dirichlet Process and n follows some distribution ρ on $\{1, 2, 3, \dots\}$. Our circular density prior (3.4.1) can be seen to take the form (3.4.6) by letting $f(u \mid \mu, n) =$

$\sum_{j=0}^{2n} \mathbb{I}_{R_{j,n}}(\mu) C_{j,n}(u)$. This point of view is especially useful in view of the Slice Sampler of Walker (2007); Kalli et al. (2011) which is tailored to Dirichlet Process Mixtures (DPMs).

Furthermore, Theorem 3.4.3 may be applied to a class of such DPMs. The idea is the following. In order to describe properties of (3.4.6), consider the linear operators T_n, $n \in \mathbb{N}$, which maps a probability measure P on $\mathbb{M}$ to the density

$$T_n P = \int_{\mathbb{M}} f(\cdot \mid \mu, n) P(d\mu). \tag{3.4.7}$$

If P has some continuous density p, then it is natural to require that $\|T_n P - p\|_\infty \xrightarrow{n\to\infty} 0$ (see e.g. assumption A2 in Bhattacharya and Dunson (2012)). If also the image under T_n of all absolutely continuous probability measures is a finite dimensional space, then Theorem 3.4.3 can be applied to ensure strong posterior consistency.

For instance, we can let

$$f(u \mid \mu, n) = C_{0,n}(u - \mu) \tag{3.4.8}$$

to obtain a Dirichlet process mixture over a continuous range of locations. The associated operator T_n defined by (3.4.7), when seen as acting on probability densities, is the De la Vallée Poussin mean of Pólya and Schoenberg (1958). Now for any density f on $\mathbb{S}^1$, $T_n f$ is a trigonometric polynomial of degree n (Pólya and Schoenberg, 1958). Hence the dimension of $T_n(\mathbb{F})$ is bounded above by $2n + 1$. Following general theory about integral operators (DeVore and Lorentz, 1993), it is straightforward to verify that $\|T_n f - f\|_\infty \to 0$ for all continuous f. Theorem 3.4.3 is therefore immediately applied to obtain strong posterior consistency.

In Section 3.5, a prior of the type (3.4.6) with densities given by (3.4.8) is compared to our circular density prior (3.4.1). Both yield very similar posterior mean estimates in our examples.

3.4.4 Adaptative convergence rates

It is interesting to note that the framework of Section 3.4.2 may be precised as to obtain adaptative convergence rates on classes of smooth densities, similarily as in Kruijer and van der Vaart (2008); Shen and Ghosal (2015). Again, the posterior convergence result is stated in some generality as to be easily applicable to other problems of similar nature.

Here we write $a_n \asymp b_n$ if there are positive constants A and B such that $Ab_n \leqslant a_n \leqslant Bb_n$ for all large n. The posterior distribution of Π is said to contract around f_0 at the rate ε_n if $(X_i)_{i\geqslant 1} \sim P_{f_0}^{(\infty)}$ implies that for all large $L > 0$,

$$\Pi\left(\{f \in \mathbb{F} : H(f_0, f) < L\varepsilon_n\} \mid (X_i)_{i=1}^n\right) \to 1, \quad P_{f_0}^{(\infty)}\text{-a.s.} \tag{3.4.9}$$

where $H(f_0, f) = \left(\int(\sqrt{f_0} - \sqrt{f})^2\right)^{1/2}$ is the Hellinger distance.

The following assumptions are made on the sequence of operators T_n and on the distribution ρ which induces the prior Π defined by (3.4.5) with Π_n priors on the submodels $T_n(\mathbb{F})$. The proof of Theorem 3.4.4 is in the appendix.

A1 The sequence of linear operators $T_n : L^1(\mathbb{M}) \to L^1(\mathbb{M})$ with $T_n(\mathbb{F}) \subset \mathbb{F}$ maps densities to densities and is such that $\|T_n 1 - 1\|_\infty \to 0$ for the constant function 1.

A2 There exists $d_n \in \mathbb{N}$ an increasing integer sequence with $d_n \geqslant \dim(T_n(\mathbb{F}))$ and satisfying $d_n \asymp n^d$ for some $d \geqslant 1$.

A3 The distribution ρ on $\mathbb{N}$ satisfies $\log(\rho(n)) \asymp -d_n \log(d_n)$.

Theorem 3.4.4. *Suppose that **A1**, **A2** and **A3** are satisfied. Let $f_0 \in \mathbb{F}$ be such that $\|\log f_0\|_\infty < \infty$, $\|T_n f_0 - f_0\|_\infty = \mathcal{O}(n^{-\beta})$ for some $\beta > 0$ and suppose there*

exists $\kappa > 0$, $\varepsilon_0 > 0$ *such that for every large* $n \in \mathbb{N}$ *and every* $0 < \varepsilon < \varepsilon_0/d_n$,

$$\Pi_n\left(\{f \in T_n(\mathbb{F}) : \|f - T_n f_0\|_\infty \leqslant \varepsilon\}\right) \geqslant (\varepsilon/d_n)^{\kappa d_n}. \tag{3.4.10}$$

Then the posterior distribution of Π *contracts around* f_0 *at the rate* $\varepsilon_n = (n/\log(n))^{-\beta/(2\beta+d)}$.

Remark 3.4.2. In order to verify (3.4.10), suppose as in (3.3.4) that

$$T_n(\mathbb{F}) = \left\{ \sum_{j=0}^{d_n} c_{j,n}\phi_{j,n} \mid (c_{j,n})_{j=0}^{d_n} \in \Delta_{d_n} \right\}$$

for some families of basis functions $\{\phi_{j,n}\}_{j=0}^{d_n}$ with $\max_j \|\phi_{j,n}\|_\infty \leqslant Cd_n$ for some $C > 0$ that does not depend on n. Writing $f = \sum_{j=0}^{d_n} c_{j,n}\phi_{j,n}$ and $T_n f_0 = \sum_{j=0}^{d_n} c_{j,n}^{(0)}\phi_{j,n}$, we find $\|f - T_n f_0\|_\infty \leqslant Cd_n \sum_{j=0}^{d_n} |c_{j,n} - c_{j,n}^{(0)}|$. Now consider a Dirichlet distribution P on the coefficients $(c_{j,n})_{j=0}^{d_n}$ with parameters $(\alpha_{j,n})_{j=0}^{d_n}$ satisfying $\sum_{j=0}^{d_n} \alpha_{j,n} = \alpha$ and $ad_n^{-1} < \alpha_{j,n} < b$ for some positive constants α, a and $b > 1$ that do not depend on n. An application of Lemma A.1 of Ghosal (2001) yields that for every $0 < \varepsilon < \min\{1, 2C/b\}$ and $d_n \geqslant 2$,

$$\begin{aligned}\Pi_n\left(\{f \in T_n(\mathbb{F}) : \|f - T_n f_0\|_\infty \leqslant \varepsilon\}\right) &\geqslant P(\{(c_{j,n})_{j=0}^{d_n} : \sum_{j=0}^{d_n} |c_{j,n} - c_{j,n}^{(0)}| \leqslant (Cd_n)^{-1}\varepsilon\}) \\ &\geqslant (\varepsilon/d_n)^{\kappa d_n}\end{aligned}$$

for some $\kappa > 0$ that does not depend on n.

Remark 3.4.3. In the case where $f_0 \in T_k(\mathbb{F})$ for some $k \in \mathbb{N}$, the use of $T_n f_0$ to control the approximation error to the sieves may be suboptimal. In this case, it is possible to obtain convergence rates of the order of $(n/\log(n))^{-1/2}$. See for instance Ghosal (2001); Kruijer and van der Vaart (2008); Barrientos et al. (2015).

Remark 3.4.4. The work in this section shares similarities to Shen and Ghosal (2015) who also obtained general adaptative contraction rates of posterior distributions for a class of random series priors. The reader is refered to Petrone and Veronese (2010) for a different generalization of the random Bernstein polynomials that is also based on constructive approximation techniques.

Application to a circular density prior

Let us continue the example of Section 3.4.3, where the prior Π on the space of all bounded circular densities is a Dirichlet Process location Mixture of $C_{0,n}$ with a distribution ρ on $n \in \mathbb{N}$. The corresponding operator T_n is defined in (3.4.7) using the densities (3.4.8). If ρ is chosen so that $\log(\rho(n)) \asymp -n\log(n)$ and the base distribution of the Dirichlet Process is uniform on $\mathbb{S}^1$ with concentration parameter $\alpha > 0$, Theorem 3.4.4 is easily applied as to obtain the rate of convergence $(n/\log(n))^{-\beta/(2\beta+2)}$ when f_0 is such that $\|\log f_0\|_\infty < \infty$ and satisfies the Hölder continuity condition

$$\sup_{x,y\in\mathbb{S}^1} \frac{|f_0(x) - f_0(y)|}{d_{\mathbb{S}^1}(x,y)^\beta} < \infty$$

for some $\beta \in (0,1]$. Indeed, the operator T_n satisfies the hypothesis **A1** of Theorem 3.4.4 and **A2-A3** have already been show to hold. Using Remark 3.4.2 and the fact that the distribution Π_n on the image of T_n corresponds to a Dirichlet distribution on the coefficients of the mixture $\sum_{j=0}^{2n} c_{j,n}C_{j,n}$ with parameters $\alpha_{j,n} = \frac{\alpha}{2n+1}$, we obtain that (3.4.10) is satisfied. Furthermore, (DeVore and Lorentz, 1993, eq. (8.6), Chapter 9) shows that $\|T_n f_0 - f_0\|_\infty = \mathcal{O}(\omega_{f_0}(n^{-1/2}))$, where ω_{f_0} is the modulus of continuity of f_0 defined as

$$\omega_{f_0}(\delta) = \sup\left\{|f_0(x) - f_0(y)| : x, y \in \mathbb{S}^1,\, d_{\mathbb{S}^1}(x,y) < \delta\right\}.$$

We thus obtain the stated convergence rate $\varepsilon_n = (n/\log(n))^{-\beta/(2\beta+2)}$ which is, up to log factors, the same as in the case of the random Bernstein polynomial prior (Kruijer and van der Vaart, 2008) for $\beta \in (0,1]$. In the case where f_0 is continuously differentiable with f_0' satisfying the Hölder continuity condition with parameter $\alpha \in (0,1]$, then (DeVore and Lorentz, 1993, eq. (8.6), Chapter 9) together with (DeVore and Lorentz, 1993, eq. (7.13), Chapter 2) shows that $\|T_n f_0 - f_0\|_\infty = \mathcal{O}(n^{-(1+\alpha)/2})$. This yields the posterior contraction rate $\varepsilon_n = (n/\log(n))^{-(1+\alpha)/(2(1+\alpha)+2)}$ which is again the same, up to log factors, as for the

random Bernstein polynomial prior (Kruijer and van der Vaart, 2008). Similar arguments may be used to obtain contraction rates in the case of the De la Vallée Poussin prior (3.4.1).

3.5 Comparison of density estimates

In this section, we compare density estimates based on the De la Vallée Poussin basis and the nonnegative trigonometric sums of Fernández-Durán (2004). Focus is on the expected Kullback-Leibler and L^1 losses in the estimation of target densities exhibiting a range of smoothness, skewness and multimodal characteristics.

3.5.1 Nonnegative trigonometric sums

Trigonometric polynomials that are probability density functions on the circle can be parameterized by the surface of a complex hypersphere (Fernández-Durán, 2004). A circular distribution of the corresponding family takes the form

$$f(u; c_0, \ldots, c_M) = \left\| \sum_{k=0}^{M} c_k e^{iku} \right\|^2, \tag{3.5.1}$$

where the coefficients c_k are complex numbers such that $\sum_{k=0}^{M} \|c_k\|^2 = \frac{1}{2\pi}$.

The parameterization (3.5.1) is exploited in Fernández-Durán (2004, 2007); Fernández-Durán and Gregorio-Domínguez (2010, 2014a,b) to model distributions of circular random variables. Circular density estimates from i.i.d. samples are obtained therein by maximum likelihood. Goodness of fit for different degrees M of the trigonometric polynomials is assessed using Akaike's information criterion (AIC) and the Bayesian information criterion (BIC). Recently, Fernández-Durán and Gregorio-Domínguez (2016a) considered a uniform prior on the coefficients c_k, with respect to hyperspherical surface measure for the Bayesian analysis of circular distributions.

3.5.2 Methods

The following five estimates of circular densities, denoted *pd*, *pc*, *nAIC*, *nBIC* and *fdbayes*, are compared.

pd: The posterior mean estimate based on the De la Vallée Poussin prior (3.4.1). This prior is parameterized by a Dirichlet process $\mathcal{D}$ and a probability distribution ρ on $\mathbb{N}$. We chose $\mathcal{D}$ to be centered on the circular uniform distribution with concentration parameter $\alpha = 1$, and we let $\rho(n) \propto e^{-n/5}$.

pc: The posterior mean estimate based on the Dirichlet process location mixture (3.4.8). This prior is also parameterized by a Dirichlet process and a distribution ρ on $\mathbb{N}$. We use the same hyperparameters as above.

nAIC: The maximum likelihood estimate of (3.5.1) where the dimension M is chosen as to minimize Akaike's information criterion.

nBIC: The maximum likelihood estimate of (3.5.1) where the dimension M is chosen as to minimize the Bayesian information criterion.

fdbayes: The posterior mean estimate based on a uniform hyperspherical distributions on the coefficients c_k of (3.5.1) and a uniform prior on $\{0, 1, 2, \ldots, 5\}$ for the dimension M. This prior on M, uniform on a range $\{0, 1, \ldots, m\}$ of values, is suggested in Fernández-Durán and Gregorio-Domínguez (2016a). The value of $m = 5$, also suggested therein, was chosen as to provide the best performance of this estimator in the comparison of Section 3.5.3.

We assess the quality of a density estimate f using the Kullback-Leibler loss defined by $\int_{\mathbb{S}^1} \log\left(\frac{f_0(u)}{f(u)}\right) f_0(u)du$, where f_0 is the target density (Kullback and Leibler, 1951), as well as the L^1 loss defined by $\int_{\mathbb{S}^1} |f_0(u) - f(u)|du$. This Kullback-

Leibler loss is appropriate in the context of discrimination between density estimates (Hall, 1987), while the L^1 loss is relevant in view of Theorem 3.4.3. Results obtained using the L^2 and Hellinger losses were highly similar to those using the L^1 loss and we omit their presentation.

Target densities

We consider the following two families of target densities to be estimated.

1. The *Skewed von Mises* family parameterized by $\alpha \in [0, 1]$ and with densities

$$v_\alpha(u) \propto (1 + \alpha \sin(u + 1)) \exp(3\alpha \cos(u - \pi)).$$

2. The family parameterized by $\alpha \in [0, 2\pi)$ and with densities

$$w_\alpha(u) \propto \exp(\sin(\cos(2u) + \sin(3u) + \alpha)),$$

which we will refer to as the w-family.

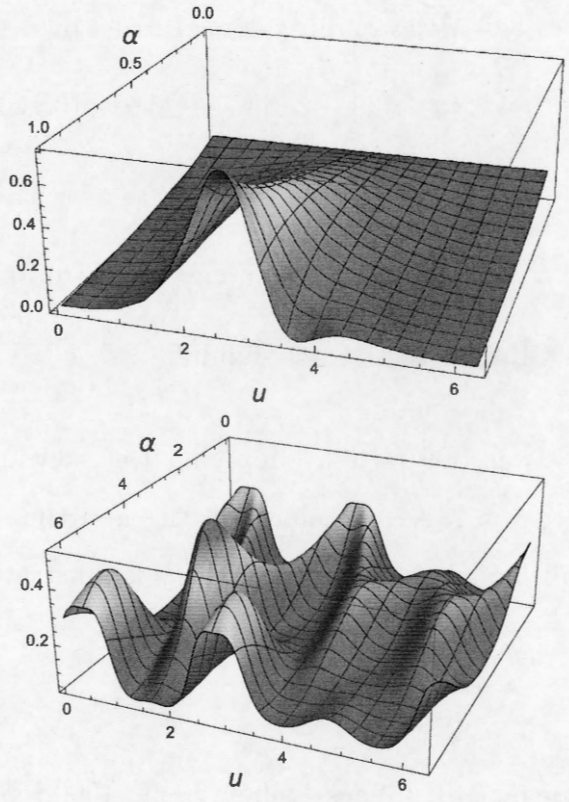
The first family was obtained by applying the skewing technique of Abe and Pewsey (2011) to von Mises circular densities and the second family was chosen to showcase multimodal characteristics. This is illustrated in Figure 3.2.

3.5.3 Results

We estimated the mean Kullback-Leibler loss in 1000 repetitions of the estimation of our target densities, for a range of parameter values, using independent samples of sizes 30 and 100. The results are shown in Figure 3.3 and Figure 3.4. Bootstrap confidence intervals at the 95% level are illustrated by vertical bars.

Under the Kullback-Leibler loss, the $nAIC$ and $nBIC$ estimators are at a considerable disadvantage in the examples considered herein. This is due to their

Figure 3.2: The *Skewed von Mises* family of densities (left panel) and the w-family of densities (right panel).



tendency of underestimating probabilities in regions where few samples are observed. An important exception to this, however, is in the use of the the *nBIC* method to estimate a constant density, since it typically selects $M = 0$ or $M = 1$ in this case and stays bounded away from zero.

The Bayesian averaging methods *pc*, *pd* and *fdbayes* are generally more appropriate under the Kullback-Leibler loss and all three are competitive. The *fdbayes* estimator has a poorer performance in the estimation of a spiked unimodal density (*Skewed von Mises* with parameter α near 1), but improves as the target density approaches being constant.

The *nAIC* estimator improves under a L^1 loss. Its increased flexibility over *nBIC* allows to better approach the target in regions of high probability density. The ordering of the estimators is otherwise roughly similar. Under a sample size of size 100, the different estimators are more clearly distinguished and the *pc* and *pd* estimators provide the best overall performance.

Remark 3.5.1. These results show that the De la Vallée Poussin densities provide a viable alternatives to the nonnegative trigonometric sums of Fernández-Durán (2004) and that they can be used to adapt techniques developped on the unit interval, such as the random Bernstein polynomials of Petrone (1999); Petrone and Wasserman (2002), to the topology of the circle. However, it is not our goal to provide best-possible estimators. It would be required to adapt the basis densities as in Kruijer and van der Vaart (2008) in order to obtain certain minimax-optimal Hellinger convergence rates. Our theoretical results can also be applied when using different density bases, including for multivariate density estimation, and the shape-preserving properties of the De la Vallée Poussin densities can be used to incorporate prior information.

Figure 3.3: Mean Kullback-Leibler losses for the *Skewed von Mises* family $\{v_\alpha\}$ of target densities and different values of the parameter α.

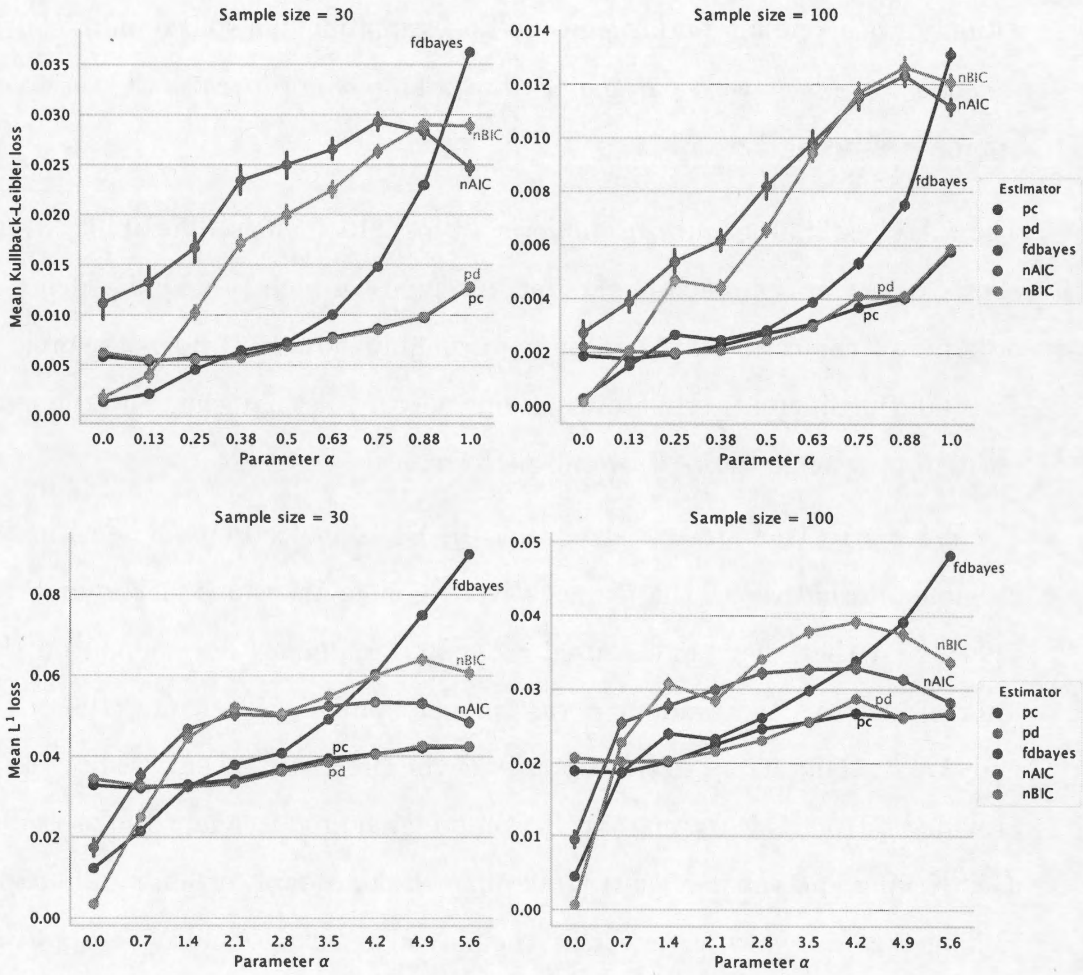


Figure 3.4: Mean Kullback-Leibler losses for the w-family $\{w_\alpha\}$ of target densities and different values of the parameter α.

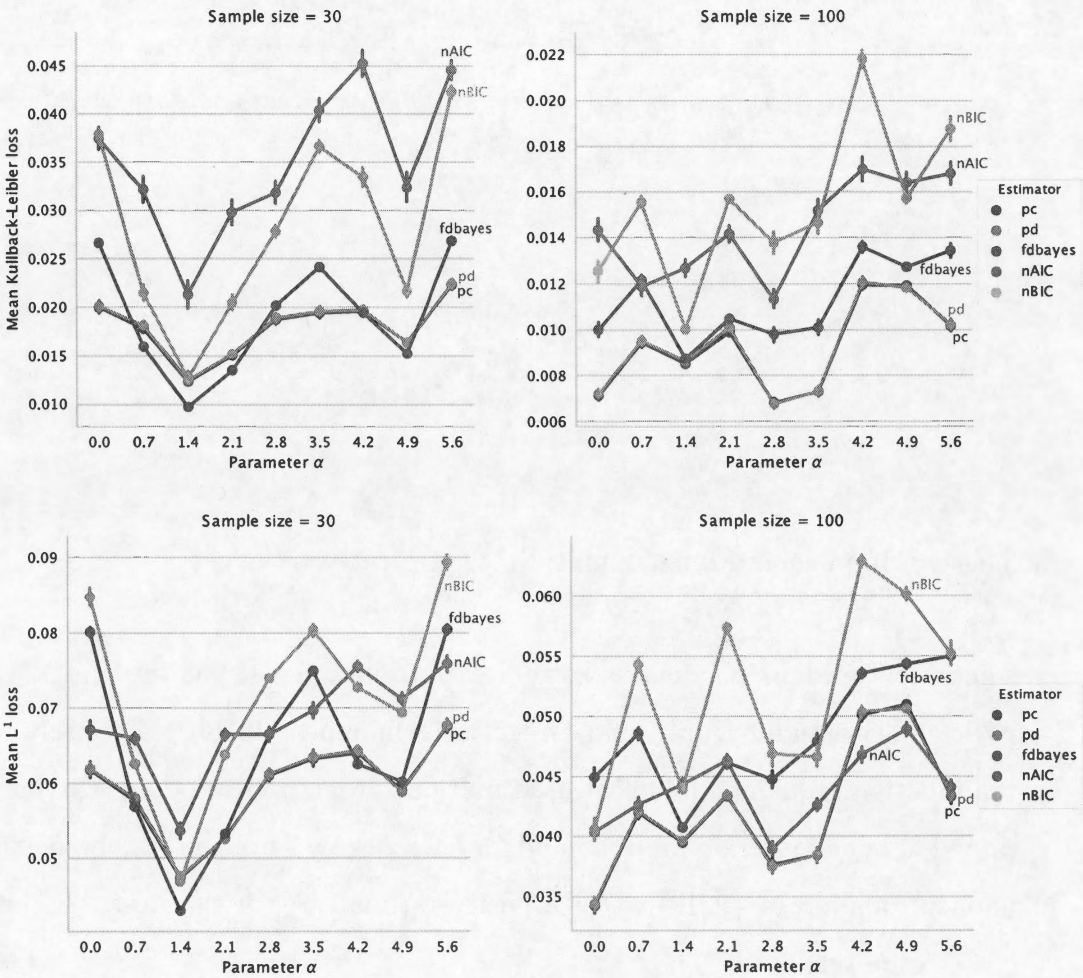
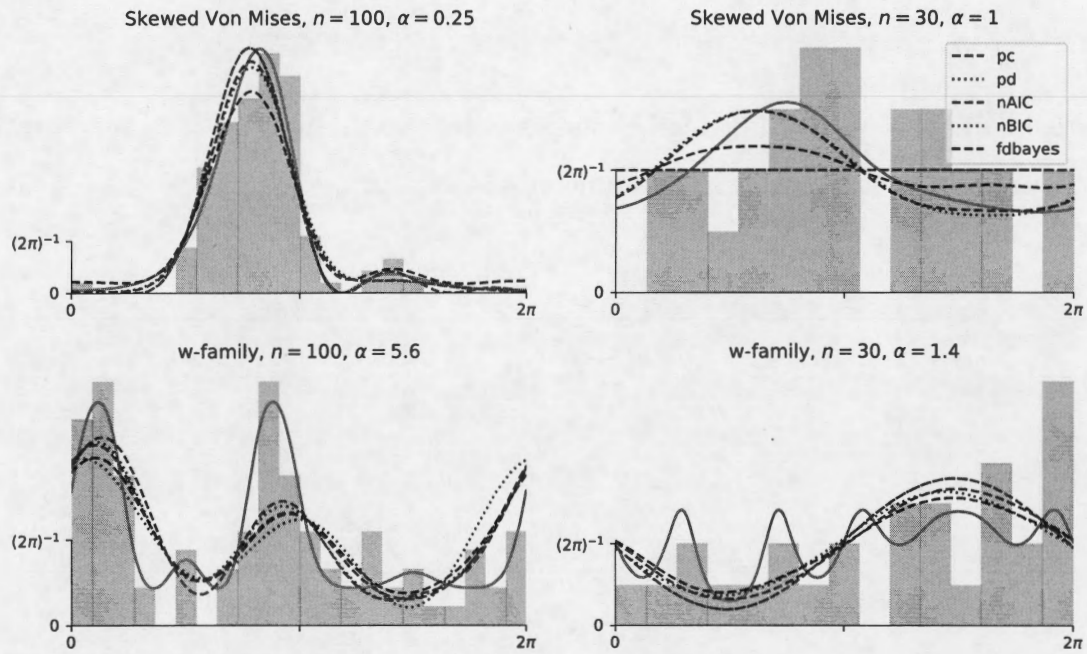


Figure 3.5: Examples of density estimates for different targets and sample sizes.



3.5.4 Implementation summary

The *nAIC* and *nBIC* density estimates are obtained using the CircNNTSR R package (Fernández-Durán and Gregorio-Domínguez, 2016b). Precisely, we ran the function "nntsmanifoldnewtonestimation" twice from random starting points provided by "nntsrandominitial" and for each degree M of the trigonometric polynomials ranging in $\{0, 1, \ldots, 7\}$. Density estimates with the best AIC and BIC scores were retrieved.

Posterior means corresponding to the *pc* and *pd* estimates are approximated using the Slice Sampler described in Kalli et al. (2011). The implementation is straightforward. We ran 80 thousand iterations of the algorithm, of which 20 thousand were treated as burn-in, and sub-sampled down to 20 thousand iterations in order

to calculate the posterior mean. Each iteration consisted in the update of every variable in the Slice Sampler following their full conditional distribution. The distribution of the model dimension n was truncated to the range $\{1, 2, 3, \ldots, 60\}$.

Posterior means for the *fdbayes* estimates are approximated using a simple independent Metropolis-Hastings algorithm with trans-dimensional moves that naturally exploit the nestedness of the models. We ran the algorithm for a million iterations, treating 100 thousand as burn-in, and sub-sampled down to 20 thousand observations in order to calculate the posterior mean. This large number of iterations was used to ensure convergence across the 7200 different datasets and to compensate for the lower acceptance rate of independent Metropolis-Hastings.

3.6 Discussion

We introduced the density basis $C_{j,n}$, $j \in \{0, 1, \ldots, 2n\}$, of the trigonometric polynomials. It is well suited to mixture modelling in the sense that different characteristics of the mixture density $f = \sum_{j=0}^{2n} c_{j,n} C_{j,n}$ can be easily related to the vector $c = (c_{0,n}, c_{1,n}, \ldots, c_{2n,n})$ of coefficients. For instance, Theorem 3.3.4 shows that f is constant if and only if c is constant; that it is periodically unimodal if c is periodically unimodal; and that the range of f is contained between $\frac{2n+1}{2\pi} \min\{c_{j,n}\}_{j=0}^{2n}$ and $\frac{2n+1}{2\pi} \max\{c_{j,n}\}_{j=0}^{2n}$. From the cyclic symmetry of the basis, it also follows that f is symmetric about 0 if the vector $(c_{n+1,n}, \ldots, c_{2n,n}, c_{0,n}, c_{1,n}, \ldots, c_{n,n})$ is symmetric about its center coefficient $c_{0,n}$. As yet another example, consider the problem of modelling a bivariate angular copula density $g : \mathbb{S}^1 \times \mathbb{S}^1 \to [0, \infty)$. Using the De la Vallée Poussin basis, we may let $g(u, v) = \sum_{i,j=0}^{2n} c_{i,j} C_{i,n}(u) C_{j,n}(v)$. The fact that g has constant marginal densities follows if the row sums and column sums of the matrix of coefficients $[c_{i,j}]_{i,j}$ are constant. On the interval $[0, 1]$, similar properties of the Bernstein polynomial densities have been exploited for

copula modelling and shape constrained regression (Guillotte and Perron, 2012; Chang et al., 2007). The De la Vallée Poussin basis may thus be used to adapt such procedures developed in the unit interval case to the topology of the circle.

Aknowledgements

The authors are grateful to the Natural Sciences and Engineering Research Council of Canada (NSERC) for a Discovery grant (S. Guillotte) as well as an Alexander Graham Bell Canada Graduate Scholarship (O. Binette).

3.7 Appendix A

3.7.1 Proof of Theorem 3.4.3

Let $\mathbb{F}$ be any space of bounded densities such that for all $f \in \mathbb{F}$, there exists $h \in \mathbb{F}$ with $\inf_x h(x) > 0$ and $\{(1-\alpha)f + \alpha h : 0 < \alpha < 1\} \subset \mathbb{F}$ (the assumption is used only at the end of the proof in *Claim 3*). We also recall the hypothesis $\mathcal{C}_n := T_n(\mathbb{F}) \subset \mathbb{F}$.

Some notations

Let $\|\cdot\|_\infty$ denote the supremum norm, let $\|\cdot\|_1$ denote the L^1-norm, and write $\mathrm{B}_1(f_0, \varepsilon) = \{f \in \mathbb{F} : \|f - f_0\|_1 < \varepsilon\}$, $\varepsilon > 0$, for an L^1-ball. For a subset $A \subset \mathbb{F}$ and $\delta > 0$, let $N(A, \delta)$ be the minimum number of L^1-balls of radius δ and centered in $\mathbb{F}$ needed to cover A. Let $\mathrm{KL}(f_0, f) = \int_{\{f_0>0\}} f_0 \log f_0/f \, d\mu$ be the Kullback-Leibler divergence between the densities f_0 and f, and denote $\mathrm{B}_{\mathrm{KL}}(f_0, \varepsilon) := \{f \in \mathbb{F} : \mathrm{KL}(f_0, f) < \varepsilon\}$. The *Kullback-Leibler support* of Π is the set of all densities f_0 such that $\Pi(\mathrm{B}_{\mathrm{KL}}(f_0, \varepsilon)) > 0$, for all $\varepsilon > 0$. Note that the $\mathfrak{B}$-measurability of $\mathrm{B}_{\mathrm{KL}}(f_0, \varepsilon)$ is shown in Barron, Schervish, and Wasserman (1999,

Lemma 11).

A result of Xing and Ranneby (2009)

Strong consistency on the Kullback-Leibler support of Π is ensured as a particular case of Xing and Ranneby (2009, Theorem 2) (see also Walker (2004); Lijoi et al. (2005)) which we state here in the following lemma (their result is stated in terms of the Hellinger distance which is topologically equivalent to the L^1-distance). The fact that $\mathbb{M}$ is a finitely measured compact metric space satisfies the conditions on $\mathbb{M}$ and $\mathbb{F}$ stated therein. Therefore, once we show that the lemma applies, all we need is to compute the Kullback-Leibler support.

Lemma 3.7.1. *Let $\mathcal{F}_n \subset \mathbb{F}$, $n \in \mathbb{N}$, be such that $\Pi(\cup_n \mathcal{F}_n) = 1$. Suppose there exists $\alpha : (0,1) \to [0,1)$ such that $\lim_{\delta \to 0} \delta/(1-\alpha(\delta)) = 0$ and*

$$\sum_{n=0}^{\infty} N(\mathcal{F}_n, \delta)^{1-\alpha(\delta)} \Pi(\mathcal{F}_n)^{\alpha(\delta)} < \infty \tag{3.7.1}$$

for every small $\delta > 0$. Then the posterior distribution of Π is strongly consistent at every density f_0 of its Kullback-Leibler support.

Application of the lemma

Denote $\overline{\mathcal{C}_n}$ the L^1-closure of $\mathcal{C}_n = T_n(\mathbb{F})$ in $\mathbb{F}$. We apply Lemma 3.7.1 with the disjoint $\mathfrak{B}$-measurable sets $\mathcal{F}_n = \overline{\mathcal{C}_n} \bigcap_{0 \leqslant k < n} \overline{\mathcal{C}_k}^c$, so that $\Pi(\cup_n \mathcal{F}_n) = \Pi(\cup_n \overline{\mathcal{C}_n}) = 1$ and $\Pi(\mathcal{F}_n) = \sum_{k \geqslant 0} \rho(k) \Pi_k(\mathcal{F}_n \cap \mathcal{C}_k) \leqslant \sum_{k \geqslant n} \rho(k)$. Let d_k be the strictly increasing integer sequence bounding $\dim(\mathcal{F}_k)$ and such that $\rho(k) < ce^{-Cd_k}$, so that we find $\sum_{k \geqslant n} \rho(k) < c \sum_{k \geqslant n} e^{-Cd_k} \leqslant c \sum_{k \geqslant d_n} e^{-Ck} \propto e^{-Cd_n}$. Moreover, from Lemma 1 of Lorentz (1966), $\mathcal{F}_n$ being of dimension at most d_n and contained in an L^1-ball of

radius 2, we have $N(\mathcal{F}_n, \delta) \leqslant (6/\delta)^{d_n}$. It follows that

$$\sum_{n=0}^{\infty} N(\mathcal{F}_n, \delta)^{1-\alpha(\delta)} \Pi(\mathcal{F}_n)^{\alpha(\delta)} \leqslant D \sum_{n=0}^{\infty} \exp\left(-d_n \left\{(1-\alpha(\delta))\log(\delta/6) + \alpha(\delta)C\right\}\right)$$

for some constant $D > 0$. Now let $\alpha(\delta) = (1-\delta)^{-\log(\delta)}$, noting that $\lim_{\delta \to 0} \alpha(\delta) = 1$ and

$$\alpha'(\delta) = \alpha(\delta)\left(\frac{\log(\delta)}{1-\delta} - \frac{\log(1-\delta)}{\delta}\right).$$

Hence, $\lim_{\delta \to 0} \delta/(1-\alpha(\delta)) = -\left(\lim_{\delta \to 0} \alpha'(\delta)\right)^{-1} = 0$. Furthermore, the series (3.7.1) converges provided$(1-\alpha(\delta))\log(\delta/6) + \alpha(\delta)C > 0$ for $\delta > 0$ sufficiently small. This is indeed the case since $\lim_{\delta \to 0} C\alpha(\delta) = C > 0$ and $\lim_{\delta \to 0}(1-\alpha(\delta))\log(\delta/6) = 0$.

The Kullback-Leibler support of Π

Let $\mathrm{KL}(\Pi)$ denote Kullback-Leibler support of Π; we show that $\mathbb{F} \subset \mathrm{KL}(\Pi)$. The proof is divided in the three following claims.

Claim 1: *For all $f \in L^1(\mathbb{M})$ we have $\|T_n f - f\|_1 \to 0$.*

To see this, the fact that T_n maps the densities of $L^1(\mathbb{M})$ to densities implies that $f \mapsto T_n f$, $f \in L^1(\mathbb{M})$, is monotone and we get $\|T_n f\|_1 \leqslant \|T_n |f|\|_1 \leqslant \|f\|_1$, for all $n \geqslant 0$. Take $\varepsilon > 0$, we can find g continuous with $\|f - g\|_1 < \varepsilon/3$; this is because the set of continuous functions on $\mathbb{M}$ is dense in $L^1(\mathbb{M})$. Now by assumption there exists $N \geqslant 0$ such that $\|T_N g - g\|_\infty < \varepsilon/(3\mu(\mathbb{M}))$, and we get $\|T_N f - f\|_1 \leqslant \|T_N(f-g)\|_1 + \|T_N g - g\|_1 + \|g - f\|_1 < \varepsilon$.

Now let $\mathbb{F}^+$ be the densities in $\mathbb{F}$ which are bounded away from zero.

Claim 2: $\mathbb{F}^+ \subset \mathrm{KL}(\Pi)$.

We show that for all $f_1 \in \mathbb{F}^+$, and for all $\varepsilon > 0$, there exists an $N \geqslant 0$ and $\delta > 0$

such that $\mathrm{B}_1(T_N f_1, \delta) \cap \mathcal{C}_N \subset \mathrm{B}_{\mathrm{KL}}(f_1, \varepsilon)$. The result will then follow from

$$\Pi(\mathrm{B}_{\mathrm{KL}}(f_1, \varepsilon)) = \sum_{k \geqslant 0} \rho(k) \Pi_k(\mathrm{B}_{\mathrm{KL}}(f_1, \varepsilon) \cap \mathcal{C}_k) \geqslant \rho(N) \Pi_N(\mathrm{B}_1(T_N f_1, \delta) \cap \mathcal{C}_N) > 0,$$

since $\rho(N) > 0$ and Π_N has support $\mathcal{C}_N$. To find such N and δ, notice that for all $f \in \mathbb{F}^+$,

$$\mathrm{KL}(f_1, f) \leqslant \|f_1/f\|_\infty \|f_1 - f\|_1 \leqslant \|f_1/f\|_\infty (\|f_1 - T_n f_1\|_1 + \|T_n f_1 - f\|_1). \quad (3.7.2)$$

Now put $0 < \inf_{x \in \mathbb{M}} f_1(x) =: m \leqslant M := \sup_{x \in \mathbb{M}} f_1(x)$. By the first claim, there exists $N \geqslant 0$ such that $\|T_n f_1 - f_1\|_1 < \frac{m}{8M}\varepsilon$, for all $n \geqslant N$. Furthermore, since $f \mapsto T_n f$ is monotone and since $\|T_n m - m\|_\infty \to 0$, we can assume N is large enough so that we also have $\inf_{x \in \mathbb{M}} T_N f_1(x) \geqslant \inf_{x \in \mathbb{M}} T_N m(x) \geqslant m/2$. Since $\mathcal{C}_N = T_N(\mathbb{F}) \subset \mathbb{F}$ and is finite dimensional, $\|\cdot\|_\infty$ is finite and equivalent to $\|\cdot\|_1$ on $\mathcal{C}_N$ and we can find $0 < \delta < \frac{m}{8M}\varepsilon$ such that $\mathrm{B}_1(T_N f_1, \delta) \cap \mathcal{C}_N \subset \mathrm{B}_\infty(T_N f_1, m/4) \cap \mathcal{C}_N$. Now for any $f \in \mathrm{B}_1(T_N f_1, \delta) \cap \mathcal{C}_N$, the quantity $\|f_1/f\|_\infty \leqslant 4M/m$, so that by plugging N in (3.7.2) we get $\mathrm{KL}(f_1, f) < \varepsilon$.

Claim 3: $\mathbb{F} \setminus \mathbb{F}^+ \subset \mathrm{KL}(\Pi)$.

Let $f_0 \in \mathbb{F} \setminus \mathbb{F}^+$ and let $0 < \varepsilon < 6$. By assumption there is an $h \in \mathbb{F}^+$ such that $\{(1-\alpha)f_0 + \alpha h : 0 < \alpha < 1\} \subset \mathbb{F}$. Now take $f_1 = \frac{f_0 + \gamma h}{1+\gamma} \in \mathbb{F}^+$, with $\gamma = \varepsilon/6$, so $f_0 < (1+\gamma) f_1$. We use the following result from Ghosal, Ghosh, and Ramamoorthi (1999, Lemma 5.1).

Lemma 3.7.2. *If f_0 and f_1 are densities with $f_0 \leqslant C f_1$, for some $C \geqslant 1$, then for any density f,*

$$\mathrm{KL}(f_0, f) \leqslant (C+1) \log C + C \left[\mathrm{KL}(f_1, f) + \sqrt{\mathrm{KL}(f_1, f)} \right].$$

Here $(2+\gamma)\log(1+\gamma) < \varepsilon/2$. By the second claim and the above lemma, there exists $\delta > 0$ and $N \geqslant 0$ such that for $f \in \mathrm{B}_1(T_N f_1, \delta) \cap \mathcal{C}_N$, we have $\mathrm{KL}(f_0, f) < \varepsilon$.

3.8 Appendix B

3.8.1 Proof of Theorem 3.4.4

We apply a particular case of (Xing, 2011, Theorem 1) which is stated in the following lemma. Here $H(f_0, f)^2 = \int \left(\sqrt{f} - \sqrt{f_0}\right)^2 d\mu$ is the squared Hellinger distance and $N(\varepsilon, \mathcal{F}; H)$ is the covering number of $\mathcal{F}$ with respect to the Hellinger distance: it is the minimum number of Hellinger balls of radius ε necessary to cover $\mathcal{F}$.

Lemma 3.8.1 (Xing (2011)). *Let ε_n and $\tilde{\varepsilon}_n$ be positive sequences such that $n \min\{\varepsilon_n^2, \tilde{\varepsilon}_n^2\} \to \infty$ as $n \to \infty$. Suppose there exists subsets $\mathcal{F}_j$, $j \in \mathbb{N}$, of $\mathbb{F}$ with $\Pi(\cup_j \mathcal{F}_j) = 1$ and constants $c_1 > 0$, $c_2 > 0$, $0 \leqslant \alpha < 1$ such that*

$$\sum_{n=1}^{\infty} e^{-c_1 n \tilde{\varepsilon}_n^2} \sum_{j=1}^{\infty} N(\tilde{\varepsilon}_n, \mathcal{F}_j; H)^{1-\alpha} \Pi(\mathcal{F}_j)^{\alpha} < \infty \tag{3.8.1}$$

and

$$\Pi\left(\left\{f \in \mathbb{F} : H(f_0, f)^2 \| f_0/f \|_{\infty}^{1/2} \leqslant \varepsilon_n^2\right\}\right) \geqslant e^{-n\varepsilon_n^2 c_2} \tag{3.8.2}$$

for all large n. Then the posterior distribution of Π contracts around f_0 at the rate $\max\{\varepsilon_n, \tilde{\varepsilon}_n\}$.

Here we let $\tilde{\varepsilon}_n = n^{-\gamma}$ for γ satisfying $\beta/(2\beta+d) < \gamma < 1/2$, and $\varepsilon_n = (n/\log(n))^{-\beta/(2\beta+d)}$. The two conditions (3.8.1) and (3.8.2) can be independently verified.

Verification of condition (3.8.1)

This follows along the lines of Section 3.1 in Xing (2008). By assumption **A3**, there exists a constant $C > 0$ such that $\rho(n) \leqslant e^{-Cd_n \log(d_n)}$. As in the proof of Theorem 3.4.3, we let $\mathcal{F}_j = \overline{\mathcal{C}_j} \bigcap_{0 \leqslant k < j} \overline{\mathcal{C}_k}^c$ with $\mathcal{C}_j = T_j(\mathbb{F})$. Now using **A2**, $\Pi(\mathcal{F}_j) \leqslant \sum_{k \geqslant j} \rho(k) \leqslant \sum_{k \geqslant d_j} e^{-Ck \log(k)}$ is bounded above by $Le^{-Cd_j \log(d_j)}$, $L = 2^C/(2^C -$

1), when $j \geqslant 2$. Since $H(f,g)^2 \leqslant \int |f-g|\, d\mu$, we have that $N(\tilde{\varepsilon}_n, \mathcal{F}_j; H) \leqslant N(\tilde{\varepsilon}_n^2, \mathcal{F}_j) \leqslant (6/\tilde{\varepsilon}_n^2)^{d_j}$ where the last inequality is derived as in Appendix 3.7.1.

Now let $0 \leqslant \alpha < 1$ be sufficiently close to 1 so that $C\alpha(1-2\gamma) \geqslant 2\gamma(1-\alpha)$. By Lemma 3.9.2, there exists $D > 0$ with $\sum_{j=1}^{\infty} \left(\frac{j^{C\alpha}}{6^{1-\alpha} n^{2\gamma(1-\alpha)}}\right)^{-j} \leqslant \exp\left(Dn^{2\gamma(1-\alpha)/(C\alpha)}\right)$ for every large n. We therefore obtain

$$\begin{aligned}
\sum_{j=1}^{\infty} N(\tilde{\varepsilon}_n, \mathcal{F}_j; H)^{1-\alpha} \Pi(\mathcal{F}_j)^{\alpha} &\leqslant L^{\alpha} \sum_{j=1}^{\infty} (6n^{2\gamma})^{d_j(1-\alpha)} e^{-Cd_j \log(d_j)\alpha} \\
&\leqslant L^{\alpha} \sum_{j=1}^{\infty} (6n^{2\gamma})^{j(1-\alpha)} e^{-Cj \log(j)\alpha} \\
&= L^{\alpha} \sum_{j=1}^{\infty} \left(\frac{j^{C\alpha}}{6^{1-\alpha} n^{2\gamma(1-\alpha)}}\right)^{-j} \leqslant L^{\alpha} \exp\left(Dn^{2\gamma(1-\alpha)/(C\alpha)}\right).
\end{aligned}$$

Taking $c_1 > D$ and since $(1-2\gamma) \geqslant 2\gamma(1-\alpha)/(C\alpha)$, it follows that

$$\begin{aligned}
&\sum_{n=1}^{\infty} e^{-n\tilde{\varepsilon}_n^2 c_1} \sum_{j=1}^{\infty} N(\tilde{\varepsilon}_n, \mathcal{F}_j)^{1-\alpha} \Pi(\mathcal{F}_j)^{\alpha} \\
&\leqslant L^{\alpha} \sum_{n=1}^{\infty} \exp\left(Dn^{2\gamma(1-\alpha)/(C\alpha)} - c_1 n^{1-2\gamma}\right) < \infty.
\end{aligned}$$

Verification of condition (3.8.2)

This follows along the lines of the proof of Theorem 2.3 in Ghosal (2001) and of the proof of Theorem 2 in Kruijer and van der Vaart (2008). Again $\varepsilon_n = (n/\log(n))^{-\beta/(2\beta+d)}$ and we let k_n be an integer sequence such that $k_n \asymp \varepsilon_n^{-1/\beta}$. The first step of the proof is to show that for some constant $L_1 > 0$ and for n sufficiently large,

$$\left\{f : H(f_0, f)^2 \|f_0/f\|_{\infty}^{1/2} \leqslant L_1 \varepsilon_n^2\right\} \supset \left\{f \in T_{k_n}(\mathbb{F}) : \|T_{k_n} f_0 - f\|_{\infty} \leqslant \varepsilon_n\right\}. \quad (3.8.3)$$

The probability of the set on the right hand side will then be lower bounded through (3.4.10).

Since $\|\log f_0\|_\infty < \infty$ by assumption, there exists constants m, M with $0 < m < f_0 < M$. Furthermore, if $f \in \mathbb{F}$ is such that $\|T_n f_0 - f\|_\infty < \inf T_n f_0$, then

$$\|f_0/f\|_\infty \leqslant \frac{M}{(\inf T_n f_0) - \|T_n f_0 - f\|_\infty}.$$

By assumption **A1** and the resulting positivity of T_n, $\inf T_n f_0 \geqslant T_n(m) \to m$ as $n \to \infty$. Hence for n sufficiently large that $\inf T_n f_0 > m/2$ and if $\|T_n f_0 - f\|_\infty < m/4$, then

$$\|f_0/f\|_\infty \leqslant \frac{M}{m/2 - \|T_n f_0 - f\|_\infty} \leqslant 4M/m.$$

Now, since we are integrating with respect to the finite measure μ, we also have

$$\begin{aligned} H(f_0, f)^2 &\leqslant \int \left(\sqrt{f} - \sqrt{f_0}\right)^2 \left(1 + \sqrt{f/f_0}\right)^2 d\mu \\ &\leqslant m^{-1} \int (f - f_0)^2 d\mu \\ &\leqslant m^{-1} \mu(\mathbb{M}) \|f - f_0\|_\infty^2. \end{aligned}$$

Furthermore, $\|f - f_0\|_\infty \leqslant \|T_{k_n} f_0 - f_0\|_\infty + \|T_{k_n} f_0 - f\|_\infty$ with $\|T_{k_n} f_0 - f_0\|_\infty = \mathcal{O}(k_n^{-1/\beta})$ and $k_n^{-\beta} \asymp \varepsilon_n$. Therefore, taking n sufficiently large that $\inf T_{k_n} f_0 > m/2$ and $\varepsilon_n \leqslant m/4$, we have that $\|T_{k_n} f_0 - f\|_\infty \leqslant \varepsilon_n$ implies

$$H(f_0, f) \|f_0/f\|_\infty^{1/4} \leqslant L_2(k_n^{-\beta} + \varepsilon_n) \leqslant L_3 \varepsilon_n$$

for some constants L_2 and L_3. This proves (3.8.3).

Now for n sufficiently large, we have $\varepsilon_n^{1+d/\beta} \leqslant \varepsilon_n$ and $\varepsilon_n^{1+d/\beta} \leqslant \varepsilon_0/d_{k_n}$, where ε_0 is a fixed constant in Theorem 3.4.4. Hence using (3.4.10) we find

$$\begin{aligned} \Pi\left(\{f \in T_{k_n}(\mathbb{F}) : \|T_{k_n} f_0 - f\|_\infty \leqslant \varepsilon_n\}\right) &\geqslant \Pi\left(\left\{f \in T_{k_n}(\mathbb{F}) : \|T_{k_n} f_0 - f\|_\infty \leqslant \varepsilon_n^{1+d/\beta}\right\}\right) \\ &\geqslant \rho(k_n) \left(\frac{\varepsilon_n^{1+d/\beta}}{d_{k_n}}\right)^{\kappa d_{k_n}}. \end{aligned}$$

Combining assumptions **A2** and **A3**, there exist positive constants A and B such that

$$\rho(k_n) \geqslant \left(\frac{1}{d_{k_n}}\right)^{A d_{k_n}} \quad \text{and} \quad d_{k_n} \leqslant B \varepsilon_n^{-d/\beta}.$$

It follows that for n sufficiently large and taking $A > \kappa$,

$$\begin{aligned}\rho(k_n)\left(\frac{\varepsilon_n^{2+d/\beta}}{d_{k_n}}\right)^{\kappa d_{k_n}} &\geqslant \left(\frac{1}{d_{k_n}}\right)^{Ad_{k_n}}\left(\frac{\varepsilon_n^{1+d/\beta}}{d_{k_n}}\right)^{\kappa d_{k_n}} \\ &\geqslant \left(\frac{\varepsilon_n^{1+2d/\beta}}{B}\right)^{AB\varepsilon_n^{-d/\beta}} \\ &\geqslant \exp\left\{-c_2 n\varepsilon_n^2\right\}\end{aligned}$$

for some positive constant $c_2 > 0$. This finishes the proof of Theorem 3.4.4.

3.9 Appendix C

3.9.1 Auxiliary results

Lemma 3.9.1. *Let μ be a finite measure on the compact metric space $(\mathbb{M}, d)$. For each $n \geqslant 0$, $d_n \geqslant 0$, let $\{\phi_{i,n}\}_{i=0}^{d_n}$ be a set of densities (with respect to μ) and let $\{R_{i,n}\}_{i=0}^{d_n}$ be a partition of $\mathbb{M}$. Let $T_n f = \sum_{i=0}^{d_n}\left(\int_{R_{i,n}} f\, d\mu\right)\phi_{i,n}$, $f \in L^1(\mathbb{M})$. If the three following conditions hold:*

(i) $\max_i \operatorname{diam}(R_{i,n}) \to 0$, *as* $n \to \infty$, *where* $\operatorname{diam}(R_{i,n}) = \sup\{d(x,y) : x, y \in R_{i,n}\}$,

(ii) for all $\delta > 0$, $\sum_{\{i:d(x,R_{i,n})\geqslant\delta\}} \mu(R_{i,n})\phi_{i,n}(x) \to 0$, *uniformly in* $x \in \mathbb{M}$, *where* $d(x, R_{i,n}) := \inf\{d(x,y) : y \in R_{i,n}\}$,

(iii) $\sum_{i=0}^{d_n} \mu(R_{i,n})\phi_{i,n} = 1$, *so that* $T_n c = c$, *for all* $c \in \mathbb{R}$,

then we have $\|T_n f - f\|_\infty \to 0$ *for every continuous density* f.

Proof. Let f be a (uniformly) continuous density on $\mathbb{M}$ and let $\varepsilon > 0$. From *(iii)* we have $|T_n f(x) - f(x)| \leqslant \sum_{i=0}^{d_n}\int_{R_{i,n}} |f(y) - f(x)|\,\mu(dy)\phi_{i,n}(x)$. Take $\varepsilon > 0$, there

exists $\delta > 0$, such that $|f(y) - f(x)| < \varepsilon/2$, for all $y \in \mathrm{B}_d(x, \delta)$. Using *(i)*, let $N \geqslant 0$ be chosen so that $\max_i \operatorname{diam}(R_{i,n}) < \delta/2$, for all $n \geqslant N$. Notice that for $n \geqslant N$, we have $\mathbb{M} = \mathrm{B}_d(x, \delta) \cup_{\{i: d(x, R_{i,n}) \geqslant \delta/2\}} R_{i,n}$; this follows from the fact that $d(x, y) \leqslant d(x, S) + \operatorname{diam}(S)$, for all $y \in S \subset \mathbb{M}$. Therefore,

$$\begin{aligned}
|T_n f(x) - f(x)| &\leqslant \sum_{i=0}^{d_n} \int_{R_{i,n}} |f(y) - f(x)| \, \mu(dy) \phi_{i,n}(x), \\
&\leqslant \frac{\varepsilon}{2} \sum_{i=0}^{d_n} \int_{R_{i,n} \cap \mathrm{B}_d(x,\delta)} \mu(dy) \phi_{i,n}(x) \\
&\qquad + 2\|f\|_\infty \sum_{\{i: d(x, R_{i,n}) \geqslant \delta/2\}} \int_{R_{i,n}} \mu(dy) \phi_{i,n}(x), \\
&< \varepsilon, \quad x \in \mathbb{M},
\end{aligned}$$

follows from *(iii)* and *(ii)* provided N is further chosen large enough. □

Lemma 3.9.2. *If $a, b \in (0, \infty)$, then as $n \to \infty$ we have*

$$\log \sum_{j=1}^{\infty} \left(\frac{j^b}{n^a} \right)^{-j} = \mathcal{O}\left(n^{a/b}\right).$$

Proof. Let $k_n = n^{\gamma/b}$ for some $\gamma > a$ and write

$$\sum_{j=1}^{\infty} \left(\frac{j^b}{n^a} \right)^{-j} \leqslant \sum_{j > k_n} \left(\frac{j^b}{n^a} \right)^{-j} + k_n \max_{1 \leqslant j \leqslant k_n} \left(\frac{j^b}{n^a} \right)^{-j}.$$

The second term on the right hand side is easily seen to be bounded by $k_n \exp\left(bn^{a/b}/e\right)$ and the first term is bounded by $\sum_{j=0}^{\infty} \left(\frac{k_n^b}{n^a} \right)^{-j} = \frac{1}{1 - n^{a-\gamma}} \xrightarrow{n \to \infty} 1$. Taking the logarithm and neglecting low order terms then yields the result. □

BIBLIOGRAPHY

Abe, T. and A. Pewsey (2011). Sine-skewed circular distributions. *Statistical Papers 52*(3), 683–707.

Agapiou, S., O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart (2017). Importance sampling: Intrinsic dimension and computational cost. *32*(3), 405–431.

Al-Lazikani, B., J. Jung, Z. Xiang, and B. Honig (2001). Protein structure prediction. *Current Opinion in Chemical Biology 5*(1), 51–56.

Aliprantis, C. D. and K. C. Border (2006). *Infinite dimensional analysis* (3 ed.). Springer.

Barrientos, A. F., A. Jara, and F. A. Quintana (2015). Bayesian density estimation for compositional data using random Bernstein polynomials. *Journal of Statistical Planning and Inference 166*, 116 – 125.

Barron, A., M. J. Schervish, and L. Wasserman (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics 27*(2), 536–561.

Basu, A., H. Shioya, and C. Park (2011). *Statistical inference: the minimum distance approach.* CRC Press.

Benedetto, J. J. and W. Czaja (2009). *Integration and modern analysis.* Birkhäuser Advanced Texts: Basler Lehrbücher. Birkhäuser Boston, Inc., Boston, MA.

Bhattacharya, A. and R. Bhattacharya (2012). *Nonparametric inference on manifolds*, Volume 2 of *Institute of Mathematical Statistics (IMS) Monographs*. Cambridge University Press, Cambridge.

Bhattacharya, A. and D. B. Dunson (2012). Strong consistency of nonparametric Bayes density estimation on compact metric spaces with applications to specific manifolds. *Annals of the Institute of Statistical Mathematics 64*(4), 687–714.

Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.

Binette, O. (2019). A note on reverse pinsker inequalities. *IEEE Transactions on Information Theory 65*(7), 4094–4096.

Binette, O. and S. Guillotte (2018). Bayesian nonparametrics for directional statistics. arXiv:1807.00305.

Böcherer, G. and B. C. Geiger (2016, Nov). Optimal quantization for distribution synthesis. *IEEE Transactions on Information Theory 62*(11), 6162–6172.

Cartwright, D. E. (1963). The use of directional spectra in studying output of a wave recorder on a moving ship. In *Ocean Wave Spectra*, pp. 203–218.

Chang, I.-S., L.-C. Chien, C. A. Hsiung, C.-C. Wen, and Y.-J. Wu (2007). Shape restricted regression with random Bernstein polynomials. *Lecture Notes-Monograph Series 54*, 187–202.

Chatterjee, S. and P. Diaconis (2018). The sample size required in importance sampling. *The Annals of Applied Probability 28*(2), 1099–1135.

Coeurjolly, J.-F. and N. Le Bihan (2012). Geodesic normal distribution on the circle. *Metrika 75*(7), 977–995.

DeVore, R. A. and G. G. Lorentz (1993). *Constructive approximation*, Volume 303 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin.

Dudley, R. M. (2002). *Real Analysis and Probability* (2 ed.). Cambridge University Press.

Fejér, L. (1916). Uber trigonometrische polynome. *Journal für die reine und angewandte Mathematik 146*, 53–82.

Feller, W. (1971). *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons, Inc., New York-London-Sydney.

Fernández-Durán, J. J. (2004). Circular distributions based on nonnegative trigonometric sums. *Biometrics 60*(2), 499–503.

Fernández-Durán, J. J. (2007). Models for circular-linear and circular-circular data constructed from circular distributions based on nonnegative trigonometric sums. *Biometrics 63*(2), 579–585.

Fernández-Durán, J. J. and M. M. Gregorio-Domínguez (2010). Maximum likelihood estimation of nonnegative trigonometric sum models using a Newton-like algorithm on manifolds. *Electronic Journal of Statistics 4*, 1402–1410.

Fernández-Durán, J. J. and M. M. Gregorio-Domínguez (2014a). Distributions for spherical data based on nonnegative trigonometric sums. *Statistical Papers 55*(4), 983–1000.

Fernández-Durán, J. J. and M. M. Gregorio-Domínguez (2014b). Modeling angles in proteins and circular genomes using multivariate angular distributions based on multiple nonnegative trigonometric sums. *Statistical applications in genetics and molecular biology 13 1*, 1–18.

Fernández-Durán, J. J. and M. M. Gregorio-Domínguez (2016a). Bayesian analysis of circular distributions based on non-negative trigonometric sums. *Journal of Statistical Computation and Simulation 86*(16), 3175–3187.

Fernández-Durán, J. J. and M. M. Gregorio-Domínguez (2016b). CircNNTSR: An R package for the statistical analysis of circular, multivariate circular, and spherical data using nonnegative trigonometric sums. *Journal of Statistical Software, Articles 70*(6), 1–19.

Ferreira, J. T. A. S., M. A. Juárez, and M. F. J. Steel (2008). Directional log-spline distributions. *Bayesian Analysis 3*(2), 297–316.

Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *The Annals of Statistics 29*(5), 1264–1280.

Ghosal, S., J. K. Ghosh, and R. Ramamoorthi (1999). Consistent semiparametric Bayesian inference about a location parameter. *Journal of Statistical Planning and Inference 77*(2), 181–193.

Ghosal, S., J. K. Ghosh, and A. W. van der Vaart (2000, 04). Convergence rates of posterior distributions. *Ann. Statist. 28*(2), 500–531.

Ghosh, J. and R. Ramamoorthi (2003a). *Bayesian Nonparametrics*. Springer-Verlag New York.

Ghosh, J. K. and R. V. Ramamoorthi (2003b). *Bayesian nonparametrics*. New York: Springer-Verlag.

Gibbs, A. L. and F. E. Su (2002). On choosing and bounding probability metrics. *International statistical review 70*(3), 419–435.

Good, I. (1985). Weight of evidence: A brief survey. In A. S. D. L. J.M. Bernardo, M.H. DeGroot (Ed.), *Bayesian Statistics 2*, pp. 249–270. North-Holland B.V.: Elsevier Science Publishers.

Guillotte, S. and F. Perron (2012). Bayesian estimation of a bivariate copula using the Jeffreys prior. *Bernoulli 18*(2), 496–519.

Guntuboyina, A. (2011). Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Transactions on Information Theory 57*(4), 2386–2399.

Hall, P. (1987). On Kullback-Leibler loss and density estimation. *The Annals of Statistics 15*(4), 1491–1519.

Halmos, P. R. and L. J. Savage (1949). Application of the radon-nikodym theorem to the theory of sufficient statistics. *20*(2), 225–241.

Hernandez-Stumpfhauser, D., F. J. Breidt, and M. J. van der Woerd (2017). The general projected normal distribution of arbitrary dimension: modeling and bayesian inference. *Bayesian Analysis 12*(1), 113–133.

Huber, P. J. (2011). Robust statistics. pp. 1248–1251. Springer.

Jammalamadaka, S. R. and A. SenGupta (2001). *Topics in circular statistics*, Volume 5 of *Series on Multivariate Analysis*. World Scientific Publishing Co., Inc., River Edge, NJ.

Kalli, M., J. E. Griffin, and S. G. Walker (2011). Slice sampling mixture models. *Statistics and Computing 21*(1), 93–105.

Kent, J. T. (1983). Identifiability of finite mixtures for directional data. *The Annals of Statistics 11*(3), 984–988.

Kruijer, W. and A. van der Vaart (2008). Posterior convergence rates for Dirichlet mixtures of beta densities. *Journal of Statistical Planning and Inference 138*(7), 1981 – 1992.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *22*(1), 79–86.

Kumar, P. and S. Chhina (2005). A symmetric information divergence measure of the csiszár's f-divergence class and its bounds. *Computers & Mathematics with Applications 49*(4), 575 – 588.

Kumar, P. and L. Hunter (2004). On an information divergence measure and information inequalities. *Carpathian Journal of Mathematics 20*(1), 51–66.

Lennox, K. P., D. B. Dahl, M. Vannucci, R. Day, and J. W. Tsai (2010). A Dirichlet process mixture of hidden Markov models for protein structure prediction. *The Annals of Applied Statistics 4*(2), 916–942.

Lennox, K. P., D. B. Dahl, M. Vannucci, and J. W. Tsai (2009). Density estimation for protein conformation angles using a bivariate von Mises distribution and Bayesian nonparametrics. *Journal of the American Statistical Association 104*(486), 586–596.

Liese, F. and I. Vajda (2006, Oct). On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory 52*(10), 4394–4412.

Lijoi, A., I. Prünster, and S. G. Walker (2005). On Consistency of nonparametric normal mixtures for Bayesian density estimation. *Journal of the American Statistical Association 100*(472), 1292–1296.

Lorentz, G. G. (1966). Metric entropy and approximation. *Bulletin of the American Mathematical Society 72*(6), 903–937.

Lorentz, G. G. (1986). *Bernstein polynomials* (Second ed.). New York: Chelsea Publishing Co.

Mardia, K. V. and P. E. Jupp (2000). *Directional statistics.* Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

McVinish, R. and K. Mengersen (2008). Semiparametric Bayesian circular statistics. *Computational Statistics & Data Analysis 52*(10), 4722–4730.

Petrone, S. (1999). Random Bernstein polynomials. *Scandinavian Journal of Statistics 26*(3), 373–393.

Petrone, S. and P. Veronese (2010). Feller operators and mixture priors in Bayesian nonparametrics. *Statistica Sinica 20*, 379–404.

Petrone, S. and L. Wasserman (2002). Consistency of Bernstein polynomial posteriors. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 64*(1), 79–100.

Pólya, G. and I. J. Schoenberg (1958). Remarks on de la Vallée Poussin means and convex conformal maps of the circle. *Pacific Journal of Mathematics 8*, 295–334.

Prokhorov, Y. V. (1956). Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications 1*(2), 157–214.

Ravindran, P. and S. K. Ghosh (2011). Bayesian analysis of circular data using wrapped distributions. *Journal of Statistical Theory and Practice 5*(4), 547–561.

Róth, Á., I. Juhász, J. Schicho, and M. Hoffmann (2009). A cyclic basis for closed curve and surface modeling. *Computer Aided Geometric Design 26*(5), 528–546.

Rudin, W. (1987). *Real and complex analysis* (Third ed.). New York: McGraw-Hill Book Co.

Sanz-Alonso, D. (2018). Importance sampling and necessary sample size: An information theory approach. *SIAM/ASA Journal on Uncertainty Quantification 6*(2), 867–879.

Sason, I. (2015). On Reverse Pinsker Inequalities. *ArXiv e-prints*.

Sason, I. and S. Verdú (2015, Oct). Upper bounds on the relative entropy and rényi divergence as a function of total variation distance for finite alphabets. In *2015 IEEE Information Theory Workshop - Fall (ITW)*, pp. 214–218.

Sason, I. and S. Verdú (2016). f -divergence inequalities. *IEEE Transactions on Information Theory 62*(11), 5973–6006.

Shen, W. and S. Ghosal (2015). Adaptive Bayesian procedures using random series priors. *Scandinavian Journal of Statistics 42*(4), 1194–1213.

Simic, S. (2009a). Best possible global bounds for jensen's inequality. *Applied Mathematics and Computation 215*(6), 2224 – 2228.

Simic, S. (2009b). Jensen's inequality and new entropy bounds. *Applied Mathematics Letters 22*(8), 1262 – 1265.

Simic, S. (2009c, Sep). On certain new inequalities in information theory. *Acta Mathematica Hungarica 124*(4), 353–361.

Simic, S. (2011). Sharp global bounds for jensen's inequality. *41*(6), 2021–2031.

Strassen, V. (1965). The existence of probability measures with given marginals. *36*(2), 423–439.

Vajda, I. (1972, Mar). On the f-divergence and singularity of probability measures. *Periodica Mathematica Hungarica 2*(1), 223–234.

Vajda, I. (2009). On metric divergences of probability measures. *Kybernetika 45*(6), 885–900.

Verdú, S. (2014). Total variation distance and the distribution of relative information. In *2014 Information Theory and Applications Workshop (ITA)*, pp. 1–3.

Walker, S. (2004). New approaches to Bayesian consistency. *The Annals of Statistics 32*(5), 2028–2043.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation 36*(1), 45–54.

Xing, Y. (2008). Convergence rates of nonparametric posterior distributions. *ArXiv e-prints*, arXiv:0804.2733.

Xing, Y. (2011). Convergence rates of nonparametric posterior distributions. *Journal of Statistical Planning and Inference 141*(11), 3382 – 3390.

Xing, Y. and B. Ranneby (2009). Sufficient conditions for Bayesian consistency. *Journal of Statistical Planning and Inference 139*(7), 2479–2489.

Yang, Y. and A. Barron (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics 27*(5), 1564–1599.