UNIVERSITA' COMMERCIALE "LUIGI BOCCONI"

PhD SCHOOL

PhD program in Statistics

Cycle: XXXII

Disciplinary Field: SECS-S/01

On Complex Dependence Structures in Bayesian Nonparametrics: a Distance–based Approach

Advisor: Antonio LIJOI

Co-Advisor: Igor PRÜNSTER

PhD Thesis by

Marta CATALANO

ID number: 3032190

Year 2021

On Complex Dependence Structures in Bayesian Nonparametrics: a Distance–based Approach

By

Marta Catalano

A thesis submitted to
Bocconi University
for the degree of
DOCTOR OF PHILOSOPHY

Advisor: Prof. Antonio Lijoi
Co–advisor: Prof. Igor Prünster

Department of Decision Sciences
Bocconi University
January 2021

Abstract

Random vectors of measures are the main building block to a major portion of Bayesian nonparametric models. The introduction of infinite–dimensional parameter spaces guarantees notable flexibility and generality to the models but makes their treatment and interpretation more demanding. To overcome these issues we seek a deep understanding of infinite–dimensional random objects and their role in modeling complex dependence structures in the data. Comparisons with baseline models play a major role in the learning process and are expressed through the introduction of suitable distances. In particular, we define a distance between the laws of random vectors of measures that builds on the Wasserstein distance and combines intuitive geometric properties with analytical tractability. This is first used to evaluate approximation errors in posterior sampling schemes and then culminates in the definition of a new principled and non model–specific measure of dependence for partial exchangeability, going beyond current measures of linear dependence. The study of dependence is complemented by the investigation of asymptotic properties for partially exchangeable mixture models from a frequentist perspective. We extend Schwartz theory to a multisample framework by relying on natural distances between vectors of densities and leverage it to find optimal contraction rates for a wide class of hierarchical models.

Acknowledgements

In the course of my work, I have benefitted of help, advice and support of many people. First of all, I would like to thank my Ph.D. advisors Antonio Lijoi and Igor Prünster for introducing me to the field of Bayesian nonparametrics and for guiding my transition from student to researcher. They approved my work and consistently supported my efforts to gain confidence and overcome obstacles in my research. I am extremely grateful for their crucial role in my personal and scientific development.
I thank Pierpaolo De Blasi for sharing his knowledge on the asymptotic properties of Bayesian models. Working with him on the last chapter of my thesis has broadened my research field in various ways.
My thanks go to the Bocconi Faculty that introduced me to fascinating topics in Statistics, with major focus on the Bayesian and subjectivist approach. Their teachings provided solid foundations to my current and future research.
The members of the BNP reading group as well as my Ph.D. colleagues freely shared new ideas and always offered scientific and personal advice: I thank them all.
Finally, I would like to thank my loved ones for helping me find my way and for their constant encouragement.

Rome - Milan November 2020

Contents

List of Symbols

$\overset{\mathrm{d}}{=}$	Equal in distribution
$\sim P$	Distributed according to $P \in \mathcal{P}_{\mathbb{X}}$
$\overset{\mathrm{iid}}{\sim} P$	Independently and identically distributed according to $P \in \mathcal{P}_{\mathbb{X}}$
$\overset{\mathrm{ind}}{\sim} P_i$	Independent and distributed according to $P_i \in \mathcal{P}_{\mathbb{X}}$
$A \setminus B$	Relative complement set
$\Rightarrow$	Weak convergence
a.s.	Almost surely
$a_n \asymp b_n$	Equal order of magnitude, i.e. $a_n b_n^{-1} \to K \neq 0$ as $n \to +\infty$
$a_n \ll b_n$	Smaller order of magnitude, i.e. $a_n b_n^{-1} \to 0$ as $n \to +\infty$
$a_n \lesssim b_n$ ($a_n \gtrsim b_n$)	Smaller (greater) up to a proportionality constant
$\alpha^{[n]}$	Ascending factorial, i.e. $\alpha^{[n]} = \alpha(\alpha+1), \ldots, (\alpha+n-1)$
$\lvert \cdot \rvert$	Cardinality of a set or absolute value of a real number
$\lVert \cdot \rVert$	Euclidean norm on $\mathbb{R}^d$
$\lVert \cdot \rVert_p$	L_p norm
$f * g$	Convolution
$\mathcal{B}(\cdot)$	Borel sets of a topological space
$\mathrm{Beta}(\alpha, \beta)$	Beta distribution
$\mathrm{Be}(c, \alpha)$	Law of beta CRM of concentration parameter $c > 0$ and base measure $\alpha \in \mathcal{M}(\mathbb{X})$
A^c	Complement set
$\mathcal{C}_\Theta$	Class of models with parameter space Θ
CRM	Completely random measure
CRV	Completely random vector
$C(X, Y)$	Fréchet class of random objects X and Y
$\mathbb{D}^{k-1}$	Projection of $\mathbb{S}^{k-1}$ on $\mathbb{R}^{k-1}$
$\mathrm{Dir}(k, \boldsymbol{\alpha})$	Dirichlet distribution of parameters $(\alpha_1, \ldots, \alpha_k)$
$\mathrm{DP}(\alpha, P_0)$	Dirichlet process with base distribution $P_0 \in \mathcal{P}_{\mathbb{X}}$ and concentration parameter $\alpha > 0$
$d_{\mathcal{W}}(\cdot, \cdot)$	Distance between CRVs built on the Wasserstein distance
$d_H(\cdot, \cdot)$	Hellinger distance
$\mathbb{E}(\cdot)$	Expected value of a random object
$\mathrm{Ga}(b, \alpha)$	Gamma CRM of rate $b > 0$ and base measure $\alpha \in \mathcal{M}(\mathbb{X})$

HDP	Hierarchical Dirichlet process
iid	Independent and identically distributed
$\mathrm{KL}(\cdot;\cdot)$	Kullback–Leibler divergence
$\mathcal{L}(X)$	Probability law of the random object X
$\mathscr{L}^+$	Lebesgue measure on the positive real axis
$\mathscr{L}_n$	Lebesgue measure on $\mathbb{R}^n$
$\log_-(\cdot)$	Negative part of the logarithm, i.e. $\max(-\log(\cdot), 0)$
$\mathcal{M}_{\mathbb{X}}$	Space of boundedly finite measures on $\mathbb{X}$
$\mathscr{M}_{\mathbb{X}}$	Borel σ–algebra on $\mathcal{M}_{\mathbb{X}}$ associated to weak$^\sharp$ topology
$\mathbb{N}$	Natural numbers
$\mathbb{N}^+$	Positive natural numbers $\mathbb{N} \setminus \{0\}$
$\mathcal{N}$	Poisson random measure
$\mathrm{NRMI}(\nu)$	Law of normalized random measure with independent increments with Lévy intensity ν
$n_\wedge$ ($n_\vee$)	$\min(n_1, \ldots, n_m)$ ($\max(n_1, \ldots, n_m)$)
n_+	$n_1 + \cdots + n_m$
$\mathscr{N}(\epsilon, T, d)$	ϵ-covering of metric space (T, d)
$\mathcal{P}_{\mathbb{X}}$	Space of probabilities on $\mathbb{X}$
$\mathscr{P}_{\mathbb{X}}$	Borel σ–algebra on $\mathcal{P}_{\mathbb{X}}$ with respect to weak convergence
$P(\cdot\|Z)$	Conditional probability with respect to the σ-algebra generated by a random object Z
$\mathbb{P}$	Probability on measure space (Ω, Σ)
$P^{(n)}$	n–fold product probability on $\mathbb{X}^n$ with marginals $P \in \mathcal{P}_{\mathbb{X}}$
$P(\phi)$	Expected value of the measurable function ϕ with respect to the probability $P \in \mathcal{P}_{\mathbb{X}}$
p	Density of the probability $P \in \mathcal{P}_{\mathbb{X}}$
$\mathrm{PRM}(\nu)$	Law of a Poisson random measure with mean measure ν
$\mathbb{R}$	Real line $(-\infty, +\infty)$
$\mathbb{R}_+$	Positive real line $[0, +\infty)$
$\mathbb{R}^2_+$	$[0, +\infty) \times [0, +\infty) \setminus \{(0, 0)\}$
$\mathbb{S}^{k-1}$	$(k-1)$-dimensional simplex
$\mathcal{S}_n$	Symmetric group of order n
$\mathrm{TV}(\cdot, \cdot)$	Total variation distance
$\mathrm{Var}(\cdot)$	Variance of a random object
$\mathcal{W}/\mathcal{W}_1/\mathcal{W}_2$	Wasserstein distances
x	Observation
$\boldsymbol{x}$	Compact notation for $(x_1, \ldots, x_k)$, for some k.
X	Observable
$\mathbb{X}$	Polish space
$\mathbb{X}^n$/ $\mathbb{X}^\infty$	n/∞-fold Cartesian product of $\mathbb{X}$
Γ	Gamma function $\Gamma(z) = \int_0^{+\infty} x^{z-1} e^{-x}\, dx$, $z > 0$
$\Gamma(a, s)$	Upper incomplete gamma function $\Gamma(a, s) = \int_s^{+\infty} e^{-t}\, t^{a-1} dt$
δ_X	Dirac measure centered at X

Θ	Space of parameters
$\tilde{\mu}$	Random measure
$\tilde{\boldsymbol{\mu}}$	Random vector of measures $(\tilde{\mu}_1, \tilde{\mu}_2)$
$\tilde{\boldsymbol{\mu}}^{\text{co}}$	Comonotonic CRV, i.e. such that $\tilde{\mu}_1 = \tilde{\mu}_2$ almost surely
Π	Prior on the space of parameters
Π_m	Probability on $\mathcal{P}_{\mathbb{X}}^m$
Σ	σ–algebra
Ω	Measure space

Introduction

The last fifty years have witnessed the rising popularity of Bayesian nonparametric models in the statistics literature. Starting from the definition of Dirichlet process by Ferguson (1973), the remarkable potential of introducing infinite–dimensional parameter spaces in the Bayesian paradigm has been recognized and widely explored. The huge increase in the flexibility and generality of the models allows to describe complex dependence structures in the data, while at the same time providing a natural quantification of uncertainty and incorporation of expert opinion. Together with the computational advances of the last decades, this brought to an increasing diffusion of Bayesian nonparametric models also in applied fields.

Of course, the greater flexibility comes at a price, making the investigation of theoretical and inferential properties of the models more demanding. This applies to the interpretation of the model, prior elicitation, posterior sampling, robustness and frequentist properties, just to mention a few aspects. To address these matters one needs a deep understanding of the inferential implications of using infinitely–dimensional random objects, typically consisting in random vectors of measures, in the specification of the model. In this thesis we pursue this goal with three leading principles: (i) measure the impact of random vectors of measures on the inferential properties of the model with numerical quantities, so to allow for meaningful comparisons between different specifications; (ii) find the analytical expression of such quantities by relying on the properties of the most common random vectors of measures, such as the independence of increments and the stick–breaking representations; (iii) gain insight on the properties of a complex model by relating it to simpler and well–known ones. In the interest of quantifying and ordering such relations ("x is more similar to z than y") we address the challenging task of introducing a distance between the laws of random vectors of measures.

We seek a distance that reflects a natural idea of similarity between distributions and at the same time is analytically tractable, so that one may fruitfully use it to interpret and tune the hyperparameters of the models. To our knowledge, this is the first work that uses distances between (vectors of) random measures in this direction. Indeed, in many cases one employs distances to understand the asymptotic behavior of a distribution, so that it suffices to provide bounds that only capture the leading behavior and typically hold up to asymptotically negligible constants. Another beaten road in more applied settings is to approximate the distance between two distributions with the one between the corresponding empirical distributions, whose convergence rates are widely

studied. In contrast, the analytical evaluation of distances is already very demanding between random vectors in $\mathbb{R}^d$. Moreover, in finite–dimensional contexts one usually relies on isometric mappings between the space of distributions and suitable spaces of measurable functions on $\mathbb{R}$, usually realized through the density function, the cumulative distribution function (cdf, typically when $d = 1$) or the characteristic function. This allows to rephrase the distance between distributions in terms of the discrepancy between measurable functions on $\mathbb{R}$, which are typically more tractable and well–studied objects. Moving from $\mathbb{R}^d$ to infinite–dimensional spaces, the task of evaluating a distance between random objects becomes even harder, since their distribution is not characterized by a measurable function in general. Indeed, the specification of the law follows more indirect schemes, usually by relating to the collection of finite–dimensional projections through uniqueness arguments, as in Kolmogorov's Extension Theorem or Ionescu–Tulcea's Theorem. Even if we manage to evaluate a distance between finite–dimensional projections, which as underlined is a difficult task on its own, it is not clear how to put the information together to obtain an overall distance between infinite–dimensional random objects. In this scenario, focusing on random vectors of measures with independent increments notably simplifies the final goal. Random measures with independent increments are typically known as completely random measures (CRMs) and thus we refer to their multi–dimensional extensions as completely random vectors (CRVs). The law of this widely used class of random objects is characterized by a deterministic measure, termed Lévy intensity, so that one may use discrepancies between measures to define a distance between the corresponding CRVs, in a similar spirit to using densities or cdfs to build distances between random variables. The main obstacle is that Lévy intensities are still rather complex objects to treat, e.g. they may have infinite mass, and, more importantly, it is not immediate to understand how a variation of the Lévy intensity impacts the distribution of the CRVs. There are two main roads that one can follow: (i) build a distance in terms of the Lévy intensities putting its interpretability on the side; (ii) define an interpretable distance based on the finite–dimensional projections and then express it in terms of the underlying Lévy intensities for concrete evaluations. Since our primary goal is to use the distance to gain insight on the model, (i) is not a flawless path. On the other hand, (ii) seems particularly demanding, since the expression of the density or cdf of the finite–dimensional projections is often not available in closed form. Nonetheless, we manage to overcome these obstacles and build an interpretable and analytically tractable distance by relying on the Wasserstein distance, whose intrinsically geometric definition makes it an ideal choice for describing the similarity between distributions. Historically, the rich structure of the Wasserstein distance has been object of study in many branches of mathematics, including transport theory, partial differential equations and ergodic theory. Recently, its attractive theoretical properties have also been supported by efficient algorithms, leading to a renewed popularity both in Statistics and in the related fields of Machine Learning and Optimization. From our perspective, the Wasserstein distance has the further benefit of providing informative comparisons between distributions with different support and without density. This property is not shared by many other common distances and divergences as the total variation distance,

the Hellinger distance or the Kullback–Leibler divergence.
With a tractable distance between CRVs at our disposal, we then deal with several intriguing problems that arise in Bayesian nonparametrics. In Chapter 2 we use it to quantify errors in approximate posterior sampling schemes for exchangeable time-to-event data. Notable Bayesian nonparametric models in this context are built on CRMs, which typically characterize both the prior and the posterior distribution. Sampling infinite–dimensional random quantities is a hard task and clever approximations are needed to provide fast and effective solutions. Nonetheless, a precise quantification of approximation errors is fundamental to guarantee their reliability, especially in terms of their impact on the final object of inference, which usually consists in the hazard rate or survival function.
The use of our distance is not confined to sampling schemes and in Chapter 3 we leverage it to create a new measure of dependence for partially exchangeable models, chief result of this thesis. Partial exchangeability is a natural generalization of exchangeability that has drawn plenty of attention in the last two decades. Dependence between different groups of exchangeable observations is flexibly modeled through the introduction of dependent random probability measures, with suitable transformations of CRVs representing a natural and popular tool to define them. The amount of prior dependence between random probabilities regulates the borrowing of information between groups and crucially impacts on the final inference and prediction. In order to measure the dependence of a model and compare different dependence structures for a principled prior elicitation, current methods consist in pairwise measures of linear dependence. More specifically, one resorts to the correlation between one–dimensional projections ($\mathrm{Cor}(\tilde{P}_1(A), \tilde{P}_2(A))$ for each Borel set A), which is clearly not ideal for objects as random probability measures. In Chapter 3 we go beyond linearity and propose a new measure of dependence in terms of closeness to the maximally dependent case, which corresponds to full exchangeability of the observations. The leading idea is to recast the problem in terms of the underlying CRV and measure the dependence as the distance between this random vector and the maximally dependent one in the same Fréchet class. This brings to the first principled and non model–specific framework for measuring dependence in partially exchangeable models. The intuitive notion of similarity stemming from the Wasserstein distance is flanked by solid analytical tools. It should be stressed that the analytical treatment of the Wasserstein distance on the Euclidean plane is a lively research topic on its own, since a general expression for the optimal coupling is currently missing, making the evaluation of the distance particularly demanding. We manage to solve this issue in our particular domain of interest and then use the Lévy intensities to evaluate tight bounds of the distance in many noteworthy models in the literature, providing a natural framework for the quantification of dependence in terms of hyperparameters.
Another interesting and not much explored topic in partially exchangeable models is the analysis of their frequentist properties, which provides an additional validation of Bayesian models. Distances are still the most natural instrument to address these topics, but with a very different flavor. Instead of comparing the laws of two infinitely–dimensional random objects, we establish convergence results towards deterministic

probability measures representing the frequentist *truth*. Thus it appears more natural to define a distance on the realizations of the random object rather than on its law, which brings to a random quantification of discrepancy. In Chapter 4 we focus our attention on partially exchangeable mixture models for multivariate density estimation for different groups of observations, where the number of groups is fixed by the design of the experiment. Our final purpose is to study the convergence rate of the random m-dimensional vector of posterior densities to the true one with respect to natural distances that combine the Hellinger distance with the ℓ_p distance in $\mathbb{R}^m$. We accomplish this task by extending Schwartz theory to partially exchangeable sequences, which is non–trivial in the realistic scenario where the observations in each group grow at different rates. To overcome this issue, the models are required to put sufficient mass around a reinforced Kullback–Leibler neighborhood of the truth. We test the efficacy of this procedure on the boosted hierarchical Dirichlet process, a generalization of the hierarchical Dirichlet process that accommodates for a faster growth of the number of discovered latent features as the sample size increases. In this way, we provide a very general framework for finding convergence rates for density estimation, which opens the way to the frequentist validation of other notable dependent priors.

This thesis consists of four almost self-contained chapters and a brief appendix, where we list some useful properties of the Wasserstein distances. After an introductory chapter, the remaining ones contain the original results of the thesis. Catalano et al. (2020) is based on the contents of Chapter 2 and Catalano et al. (2020+) on the ones of Chapter 3. Chapter 4 has inspired a forthcoming work with Pierpaolo De Blasi, Antonio Lijoi and Igor Prünster.

Chapter 1 introduces the theoretical framework of this thesis. We review some key concepts in Bayesian nonparametrics, with particular emphasis on exchangeability, partial exchangeability and completely random measures. The scope of this chapter is to fix the notation and provide a common ground whom we will refer to in further chapters.

In Chapter 2 we look into Bayesian nonparametric models for exchangeable time-to-event data. In particular, we focus on priors for the hazard rate function, a popular choice being the kernel mixture with respect to a gamma random measure. Sampling schemes are usually based on approximations of the underlying random measure, both *a priori* and conditionally on the data. The main goal we pursue is the quantification of approximation errors through the Wasserstein distance. Though easy to simulate, the Wasserstein distance is generally difficult to evaluate, strengthening the need for tractable and informative bounds. Here we accomplish this task on the wider class of completely random measures and specialize our results to the gamma random measure and the related kernel mixtures. The techniques that we introduce yield upper and lower bounds for the Wasserstein distance between hazard rates, cumulative hazard rates and survival functions.

In Chapter 3 we go beyond the exchangeability assumption. The proposal and study of dependent Bayesian nonparametric models has been one of the most active research lines in the last two decades, with random vectors of measures representing a natural and popular tool to define them. Nonetheless a principled approach to understand and quantify the associated dependence structure is still missing. In this chapter we devise a general, and non model-specific, framework to achieve this task for random measure based models, which consists in: (a) quantify dependence of a random vector of probabilities in terms of closeness to exchangeability, which corresponds to the maximally dependent coupling in the same Fréchet class, i.e. the comonotonic vector; (b) recast the problem in terms of the underlying random measures (in the same Fréchet class) and quantify the closeness to comonotonicity; (c) define a distance based on the Wasserstein metric, which is ideally suited for spaces of measures, to measure the dependence in a principled way. Several results, which represent the very first in the area, are obtained. In particular, useful bounds in terms of the underlying Lévy intensities are derived relying on compound Poisson approximations. These are then specialized to popular models in the Bayesian literature leading to interesting insights.

In Chapter 4 we stay in the domain of partial exchangeability and focus on infinite–dimensional priors that place full support on absolutely continuous probabilities with respect to a common measure, typically coinciding with the Lebesgue measure on $\mathbb{R}^d$. Despite these models being widely used in the Bayesian community, there are still many open questions in terms of frequentist validation. In particular, we study the posterior contraction rates for data distributions that are modeled as mixtures over the boosted hierarchical Dirichlet process, a generalization of the hierarchical Dirichlet process that accommodates for a faster growth of the number of discovered latent features as the sample size increases. By extending Schwartz theory to partially exchangeable sequences we uncover the interesting behavior that posterior contraction rates depend on the relation between the numbers of observations in different groups. If these are equal or at most related in a polynomial fashion, we recover the minmax rates up to a logarithmic factor. As the relation becomes exponential, the rates may deteriorate drastically.

Chapter 1

Random measures in Bayesian nonparametrics

The aim of this first chapter is to fix the notation and introduce the reader to some common instruments that will be used in further chapters. We point out that what follows does not aim at being an exhaustive introduction to the field of Bayesian nonparametric statistics nor its scopes. The point of view is the one that was found to introduce at best the main contents of the thesis.

1.1 The Bayes–Laplace paradigm

Any statistical analysis begins with a collection of data $x_1, \dots, x_n$ on some space $\mathbb{X}$. Whether one wants to take an informed decision or provide a deeper insight on their nature, $x_1, \dots, x_n$ are treated as instances (*observations*) of random variables $X_1, \dots, X_n$ (*observables*). To this end, one typically endows $\mathbb{X}$ with a Borel σ–algebra $\mathcal{B}(\mathbb{X})$ and introduces a random vector $(X_1, \dots, X_n) : (\Omega, \Sigma) \to (\mathbb{X}^n, \mathcal{B}(\mathbb{X}^n))$, where $(\Omega, \Sigma, \mathbb{P})$ is some unknown probability space and $(\mathbb{X}^n, \mathcal{B}(\mathbb{X}^n))$ is the n-fold product space with corresponding product σ-algebra. We denote by P_n the unknown distribution of the random vector, i.e. $(X_1, \dots, X_n) \sim P_n \in \mathcal{P}_{\mathbb{X}^n}$, the space of probabilities on $\mathbb{X}^n$. The statistical analysis then proceeds in two directions: the modeling consists in specifying a class $\mathcal{C}_\Theta = \{P_{n,\theta} : \theta \in \Theta\} \subset \mathcal{P}_{\mathbb{X}^n}$ which is considered a reasonable container for the unknown distribution P_n; the inference consists in choosing and implementing a criterion to select the element $P_{n,\theta}$ of $\mathcal{C}_\Theta$ that "best" represents P_n. We refer to the corresponding θ as parameter of the model and underline how the parameter space Θ may always be embedded in a subset of $\mathcal{P}_{\mathbb{X}^n}$.
In this thesis we focus on the Bayes–Laplace paradigm as a criterion for model selection. We refer to Regazzini (1996) for an exhaustive exposition of this inferential framework, which we only briefly describe. From now on we assume that both Θ and $\mathbb{X}$ are Polish spaces. The Bayesian picks a probability Π on the parameter space Θ, the *prior* distribution of the parameter, and assumes that $\theta \sim \Pi$. If every $P_{n,\theta} \in C_\Theta$ is a regular conditional probability for $X_1, \dots, X_n|\theta$, Π and C_Θ may be used to define a probability

P on $\Theta \times \mathbb{X}^n$ as follows:

$$P(A \times B) = \int_A P_{n,\theta}(B)\, d\Pi(\theta),$$

for any $A \subset \Theta$ and $B \subset \mathbb{X}^n$ measurable sets, so that $(\theta, X_1, \ldots, X_n) \sim P$. The Bayesian then performs statistical inference on the parameter by providing a version of the conditional probability $P(\theta \in \cdot\, | X_1, \ldots, X_n)$, the so–called *posterior* distribution, evaluated on the observed values $x_1, \ldots, x_n$. This entails that the Bayesian paradigm does not provide a value for the parameter, but rather an entire distribution, allowing for the evaluation of uncertainty around the estimate for θ. Moreover, when the sequence of observables is judged infinitely extendible, a class of consistent sequences $\{P_{n,\theta}\}_{n\geq 1}$ of regular conditional probabilities, together with a prior Π for the parameter, may be used to uniquely define a probability for $(X_n)_{n\geq 1} \in \mathbb{X}^\infty$. Indeed by Kolmogorov's Extension Theorem, it suffices to set for every $n \in \mathbb{N}^+$ and every measurable $A \subset \mathbb{X}^n$,

$$\mathbb{P}((X_1, \ldots, X_n) \in A) = \int_\Theta P_{n,\theta}(A)\, d\Pi(\theta).$$

It should be mentioned that some statisticians find it unorthodox to express a prior opinion on a latent quantity. In the predictive approach to inference, on the other hand, the interest shifts from the parameter θ to the *predictive distribution* $\mathbb{P}(X_{n+1} \in \cdot | X_1, \ldots, X_n)$. Indeed, by Ionescu Tulcea's Theorem, a sequence of predictive distributions is sufficient to assign a law on $(X_n)_{n\geq 1}$, so that in principle the predictive approach allows one not to specify a prior opinion for the parameter and to only focus on observable quantities.
In this thesis we will focus on sequences that are judged infinitely extendible, so that $(X_1, \ldots, X_n)$ is considered as the projection on the first n coordinates of a sequence $(X_i)_{i\geq 1}$ defined on $\mathbb{X}^\infty$. Most of our work stems from the notion of exchangeability, an intuitive symmetry assumption on the distribution of the observables.

1.2 Exchangeability and de Finetti's Theorem

The notion of exchangeability is a milestone in the probabilistic foundations of Bayesian nonparametrics. Let $\mathcal{S}_n$ denote the finite symmetric group of order n, i.e. the group of permutations of n objects. We point out that $\mathcal{S}_n$ acts in a natural way of $\mathbb{N}$, by permuting the first n numbers and acting as the identity on the remaining.

Definition 1. A sequence of random variables $(X_i)_{i\geq 1}$ is said to be *exchangeable* if for every n and every $\sigma \in \mathcal{S}_n$,

$$(X_i)_{i\geq 1} \overset{\mathrm{d}}{=} (X_{\sigma(i)})_{i\geq 1},$$

where $\overset{\mathrm{d}}{=}$ denotes the equality in distribution.

In a statistical analysis, the generality of the model and the feasibility of the inference are typically complementary aspects. Thanks to de Finetti's Representation Theorem

(de Finetti, 1937), exchangeability stands out as a positive counterexample, combining an intuitive assumption on the observables with tractability in the inference. In order to state de Finetti's Theorem, we recall that if $\mathbb{X}$ is a Polish space, the space of probabilities $\mathcal{P}_{\mathbb{X}}$ is Polish as well with respect to the weak convergence. In particular, we endow $\mathcal{P}_{\mathbb{X}}$ with the corresponding Borel σ-algebra $\mathscr{P}_{\mathbb{X}}$.

Theorem 1 (de Finetti's Representation Theorem)**.** *Let $\mathbb{X}$ be a Polish space and let $(X_i)_{i\geq 1}$ be an exchangeable sequence on $\mathbb{X}^\infty$. Then there exists a probability measure Π on $\mathcal{P}_{\mathbb{X}}$ such that, for every $n \in \mathbb{N}$ and every $A_1, \dots, A_n$ measurable subsets of $\mathbb{X}$,*

$$\mathbb{P}(X_i \in A_i \text{ for } i = 1, \dots, n) = \int_{\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{n} P(A_i)\, d\Pi(P). \tag{1.1}$$

de Finetti's Theorem provides a comprehensive way to specify the law of an exchangeable sequence through the corresponding probability Π, which is usually referred to as the de Finetti measure of the sequence. Many statistical analyses rephrase (1.1) in terms of conditional independence with respect to a random probability $\tilde{P} : \Omega \to \mathcal{P}_{\mathbb{X}}$:

$$X_1, \dots, X_n | \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}; \qquad \tilde{P} \sim \Pi.$$

It follows that under the single assumption of exchangeability the parameter space Θ reduces from a subset of $\mathcal{P}_{\mathbb{X}^n}$ to a subset of $\mathcal{P}_{\mathbb{X}}$. In the next section we discuss some common specifications for the de Finetti measure Π.

1.3 The Dirichlet process

In order to perform Bayesian inference for an exchangeable sequence of observations, thanks to de Finetti's Theorem it suffices to specify a prior distribution on $\mathcal{P}_{\mathbb{X}}$, and then find the posterior distribution through conditional probability. Until the 70's the treatment of priors with large support on $\mathcal{P}_{\mathbb{X}}$ was considered unfeasible and statisticians typically resorted to parametric models, i.e. such that $\Theta \subset \mathbb{R}^d$. Then, Ferguson (1973) came up with the Dirichlet process, a probability distribution with full weak support on $\mathcal{P}_{\mathbb{X}}$ that initiated the branch of Bayesian nonparametrics and still underlies most current developments in the field.

1.3.1 Definition

In order to introduce the Dirichlet process, we first recall the definition of its finite–dimensional counterpart, the Dirichlet distribution on the simplex. For $k \in \mathbb{N} \setminus \{0, 1\}$, let $\mathbb{S}^{k-1} = \{(x_1, \dots, x_k) : \min_i x_i \geq 0, \sum_{i=1}^k x_i = 1\} \subset \mathbb{R}^k$ be the $(k-1)$–dimensional simplex and let $\alpha_1, \dots, \alpha_k > 0$. The Dirichlet distribution $\text{Dir}(k, \boldsymbol{\alpha})$ of order k and parameter $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k)$ is a probability distribution on $\mathbb{S}_k$ defined by the following density with respect to the Lebesgue measure $\mathscr{L}(\mathbb{R}^k)$:

$$f_{k,\boldsymbol{\alpha}}(x_1, \dots, x_k) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} x_1^{\alpha_1} \dots x_k^{\alpha_k} \mathbb{1}_{\mathbb{S}^{k-1}}(x_1, \dots, x_k),$$

where Γ is the gamma function. Alternatively, if $(X_1, \ldots, X_k) \sim \text{Dir}(k, \boldsymbol{\alpha})$, one can fix $X_k = 1 - \sum_{i=1}^{k-1} X_i$ and let $(X_1, \ldots, X_{k-1})$ have the following density with respect to $\mathscr{L}(\mathbb{R}^{k-1})$:

$$g_{k,\boldsymbol{\alpha}}(x_1, \ldots, x_{k-1}) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} x_1^{\alpha_1} \ldots x_{k-1}^{\alpha_{k-1}} \left(1 - \sum_{i=1}^{k-1} x_{k-1}\right)^{\alpha_k} \mathbb{1}_{\mathbb{D}_{k-1}}(\boldsymbol{x}),$$

where $\mathbb{D}_{k-1} = \{(x_1, \ldots, x_{k-1}) : \min_i x_i \geq 0, \sum_{i=1}^{k-1} x_i \leq 1\} \subset \mathbb{R}^{k-1}$ and $\boldsymbol{x} = (x_1, \ldots, x_{k-1})$. Let now $(\Omega, \Sigma, \mathbb{P})$ be a measure space endowed with a σ–algebra Σ and a probability measure $\mathbb{P}$. A random probability on a Polish space $\mathbb{X}$ is a measurable $\tilde{P} : (\Omega, \Sigma, \mathbb{P}) \to (\mathcal{P}_{\mathbb{X}}, \mathscr{P}_{\mathbb{X}})$. In order to specify the law for a random probability measure, it may be helpful to express it in terms of a stochastic process with index set $\mathcal{B}(\mathbb{X})$, the Borel sets of $\mathbb{X}$. By Kolmogorov's Extension Theorem, the probability law of $(\tilde{P}(A))_{A \in \mathcal{B}(\mathbb{X})}$ is uniquely determined by the law it attains on any finite collection of non–intersecting sets.

Definition 2. Let $\alpha > 0$ be a concentration parameter and $P_0 \in \mathcal{P}_{\mathbb{X}}$ be a base probability measure. The *Dirichlet process* with base measure αP_0 is the only probability law on $\mathcal{P}_{\mathbb{X}}$ that satisfies

$$(\tilde{P}(A_1), \ldots, \tilde{P}(A_n)) \stackrel{\text{d}}{=} \text{Dir}(n+1, (\alpha P_0(A_1), \ldots, \alpha P_0(A_n), \alpha P_0(A^*))),$$

for any collection of non–intersecting measurable sets $A_1, \ldots, A_n$ such that $A^* = \mathbb{X} \setminus \cup_{i=1}^{n} A_i$. We write $\tilde{P} \sim \text{DP}(\alpha, P_0)$.

In order to give an interpretation to the parameters, we point out that if $\tilde{P} \sim \text{DP}(\alpha, P_0)$, then $\tilde{P}(A) \sim \text{Beta}(\alpha P_0(A), \alpha P_0(A^c))$ for any measurable set A with complementary A^c in $\mathbb{X}$. It follows that

$$\mathbb{E}(\tilde{P}(A)) = P_0(A); \qquad \text{Var}(\tilde{P}(A)) = \frac{P_0(A)(1 - P_0(A))}{\alpha + 1}.$$

We may thus regard the base probability P_0 as the mean of the process and α as the concentration of the distribution around P_0. These interpretations play a fundamental role in the prior elicitation when we use the Dirichlet process as de Finetti measure for an exchangeable sequence of observations.

1.3.2 Properties

In this section we recall some fundamental features of the Dirichlet process in terms of posterior conjugacy, predictive distribution and exchangeable partition probability function. Together with the definition of Dirichlet process, Ferguson (1973) also provided a closed form expression for the corresponding posterior distribution, which is conjugate in the following sense.

Theorem 2. *Let $\mathbb{X}$ be a Polish space and let $(X_n)_{n\geq 1}$ be an exchangeable sequence on $\mathbb{X}^\infty$ with de Finetti measure $\mathrm{DP}(\alpha, \mathrm{P}_0)$, so that $X_1, \ldots, X_n | \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}$; $\tilde{P} \sim \mathrm{DP}(\alpha, \mathrm{P}_0)$. Then the posterior distribution $\tilde{P} | X_1, \ldots, X_n \sim \mathrm{DP}(\alpha + \mathrm{n}, \mathrm{P}_0^*)$, where*

$$P_0^* = \frac{\alpha}{\alpha + n} P_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{X_i}, \tag{1.2}$$

with δ_X the Dirac measure centered at X.

The previous theorem sheds light on some key properties of the use of the Dirichlet process in statistics: as the number of observation increases, the mean estimate is a weighted sum of the prior opinion and the empirical distribution. In the meantime, the confidence on the mean estimate increases and, as the number of observations goes to infinity, the posterior distribution contracts to the empirical process.
Thanks to the work of Blackwell & MacQueen (1973), we also have a characterization of the predictive distribution, which is often addressed as generalized Pólya urn scheme or Blackwell-MacQueen urn scheme.

Theorem 3. *Let $\mathbb{X}$ be a Polish space and let $(X_n)_{n\geq 1}$ be an exchangeable sequence on $\mathbb{X}^\infty$ with de Finetti measure $\mathrm{DP}(\alpha, \mathrm{P}_0)$, so that $X_1, \ldots, X_n | \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}$; $\tilde{P} \sim \mathrm{DP}(\alpha, \mathrm{P}_0)$. Then P_0^* defined in (1.2) is the predictive distribution for $X_{n+1} | X_1, \ldots, X_n$.*

The expression of P_0^* highlights the almost–sure existence of ties between the observables $X_1, \ldots, X_n$, due to the almost–sure discreteness of the Dirichlet process. Namely if $\Pi = \mathrm{DP}(\alpha, P_0)$,

$$\Pi(P \in \mathcal{P}_{\mathbb{X}} : P \text{ is discrete}) = 1.$$

One may use the ties between $X_1, \ldots, X_n$ to obtain a random partition of the first n numbers $[n] = \{1, \ldots, n\}$, which plays a particularly important role in Bayesian nonparametric clustering methods. At the same time, the exchangeable partition probability function (Pitman, 1995) provides a simple expression for the predictive distribution, which is often exploited in sampling algorithms. We give a very brief introduction to these topics and refer to Pitman (2006) and Lijoi & Prünster (2010) for a complete treatment of random partitions and their uses in Bayesian nonparametrics. We recall that a partition of $[n]$ is an unordered collection of disjoint non–empty sets $C_1, \ldots, C_k \subset [n]$ such that $[n] = \cup_{i=1}^k C_k$. A random partition of $[n]$ is a random object on the set of all partitions of $[n]$. Given a sequence of random variables $\{X_i\}_{i \leq n}$, one can define a random partition of $[n]$ by considering the random equivalence classes $\tilde{C}_1, \ldots, \tilde{C}_k$ of the relation

$$i \sim j \text{ if and only if } X_i = X_j.$$

Let $n_1, \ldots, n_k \in \mathbb{N}^+$ such that $n_1 + \cdots + n_k = n$. We define the exchangeable partition probability function (EPPF) of order n of an exchangeable sequence $(X_n)_{n\geq 1}$ as

$$p^{(n)}(n_1, \ldots, n_k) = \mathbb{P}(|\tilde{C}_1| = n_1, \ldots, |\tilde{C}_k| = n_k),$$

where $|\cdot|$ indicates the cardinality of a set. When dealing with the Dirichlet process, the EPPF comes in a very simple form and it coincides with Ewen's sampling formula.

Theorem 4. *Let $\mathbb{X}$ be a Polish space and let $(X_n)_{n\geq 1}$ be an exchangeable sequence on $\mathbb{X}^\infty$ with de Finetti measure $\mathrm{DP}(\alpha, \mathrm{P}_0)$, so that $X_1, \ldots, X_n | \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}$; $\tilde{P} \sim \mathrm{DP}(\alpha, \mathrm{P}_0)$. Then*

$$p^{(n)}(n_1, \ldots, n_k) = \frac{\alpha^k}{(\alpha)_n} \prod_{i=1}^{k} (n_i - 1)!,$$

where $(\alpha)_n = \alpha(\alpha+1)\ldots(\alpha+n-1)$.

1.3.3 Extensions

Though the Dirichlet process provides a very tractable infinite–dimensional prior, there may be scenarios that require more flexibility in the induced posterior, predictive or clustering distribution. Consequently, there have been extensive efforts to generalize the Dirichlet process in order to obtain a wider range of updating schemes. The starting points are typically the two most notable representations of the Dirichlet process, namely as stick–breaking process and as normalized random measure. In this section we briefly describe them.

We have introduced the Dirichlet process starting from its finite–dimensional distributions. An alternative compelling way to specify its law is through the *residual allocation model* or *stick–breaking* representation, which highlights the almost surely discrete nature of the process and turns out to be very convenient for sampling schemes. Let $V_i \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha)$ and define $W_i = V_i \prod_{j=1}^{i-1} V_j$. If $Z_i \overset{\text{iid}}{\sim} P_0$ independently from $(V_i)_{i\geq 1}$, then Sethuraman (1994) proved that

$$\sum_{i=1}^{+\infty} W_i \delta_{Z_i} \sim \mathrm{DP}(\alpha, P_0).$$

Starting from this representation, one possible way of extending the Dirichlet process is to assign different laws to the weights $(W_i)_{i\geq 1}$ and to the atoms $(Z_i)_{i\geq 1}$. Following Pitman (1995), one usually addresses as *species sampling models* those probability laws whose atoms are iid and independent from the weights. A prominent prior in this class is the Pitman–Yor process or two parameters Dirichlet process (Perman et al., 1992; Pitman & Yor, 1997).

Definition 3. Let $\alpha \in [0, 1)$, $\theta > -\alpha$ and $P_0 \in \mathcal{P}_{\mathbb{X}}$. Let $V_i \overset{\text{ind}}{\sim} \mathrm{Beta}(1-\alpha, \theta + i\alpha)$, $W_i = V_i \prod_{j=1}^{i-1} V_j$ and $Z_i \overset{\text{iid}}{\sim} P_0$ independently from $(V_i)_{i\geq 1}$. Then the *Pitman–Yor process* is defined by

$$\sum_{i=1}^{+\infty} W_i \delta_{Z_i} \sim \mathrm{PY}(\alpha, \theta, P_0).$$

When $\alpha = 0$ one retrieves the Dirichlet process $\mathrm{DP}(\theta, P_0)$ as special case.

As already mentioned, the Dirichlet process may also be represented as the normalization of a gamma random measure (Ferguson, 1973). This representation proves to be particularly convenient for analytical computations and offers a nourishing ground

for many generalizations of the Dirichlet process. In order to properly define gamma random measures and their extensions, in the next section we give a brief account on completely random measures.

1.4 Completely random measures

Completely random measures are one of the main building blocks for defining infinite–dimensional priors and performing Bayesian inference. In this section we revise some key properties and dwell on those aspects that will play a dominant role in future sections. For a complete account on the topic we refer to Lijoi & Prünster (2010).

1.4.1 Definition

Let $\mathcal{M}_{\mathbb{X}}$ denote the space of boundedly finite measures on $\mathbb{X}$ endowed with the weak$^\sharp$ topology, so that $\mathcal{M}_{\mathbb{X}}$ is a Polish space (Daley & Vere-Jones, 2002). We denote by $\mathscr{M}_{\mathbb{X}}$ the corresponding Borel σ-algebra. A random measure is a measurable function from some probability space $(\Omega, \Sigma, \mathbb{P})$ to $(\mathcal{M}_{\mathbb{X}}, \mathscr{M}_{\mathbb{X}})$.

Definition 4. A random measure $\tilde{\mu}$ is a *completely random measure* (CRM) if, given a finite collection of pairwise disjoint bounded measurable sets $\{A_1, \cdots, A_n\}$ of $\mathbb{X}$, the random variables $\tilde{\mu}(A_1), \cdots, \tilde{\mu}(A_n)$ are mutually independent.

An important class of CRMs is given by Poisson random measures, which are the main building block to all other CRMs. Let ν be a measure on $\mathbb{X}$. A CRM $\tilde{\mu}$ is a *Poisson random measure* of mean measure ν if $\tilde{\mu}(B)$ has a Poisson distribution of parameter $\nu(B)$, for every measurable set B such that $\nu(B) < +\infty$. We write $\tilde{\mu} \sim \text{PRM}(\nu)$ and point out that $\nu(B) = \mathbb{E}(\mathcal{N}(B))$. Kingman (1967) proved that any CRM can be uniquely represented as the sum of three independent components, $\mu + \tilde{\mu}_f + \tilde{\mu}$, where μ is a fixed measure on $\mathbb{X}$, $\tilde{\mu}_f$ is a discrete random measure with fixed atoms and $\tilde{\mu}$ is an almost surely discrete random measure without fixed atoms. In what follows we will focus on CRMs $\tilde{\mu}$ without fixed atoms, so that their distribution is uniquely determined by a Poisson random measure. Indeed, Kingman (1967) proved that there exists $\mathcal{N} \sim \text{PRM}(\nu)$ such that

$$\tilde{\mu}(dy) \overset{\text{d}}{=} \int_{\mathbb{R}^+} s\, \mathcal{N}(ds, dy), \tag{1.3}$$

where the mean measure ν on $\mathbb{R}^+ \times \mathbb{X}$ is such that for all $x \in \mathbb{X}$, $\nu(\mathbb{R}^+ \times \{x\}) = 0$, and for all bounded measurable set $A \subset \mathbb{X}$ and $\epsilon > 0$,

$$\int_A \int_{\mathbb{R}^+} \min(\epsilon, s)\, \nu(ds, dy) < +\infty. \tag{1.4}$$

We write $\tilde{\mu} \sim \text{CRM}(\nu)$ and refer to ν as the *Lévy intensity* of $\tilde{\mu}$. In view of (1.3), the probability distribution of $\tilde{\mu}$ can be characterized through the Laplace functional

$$\mathbb{E}\Big(e^{-\int_{\mathbb{X}} f(y)\tilde{\mu}(dy)}\Big) = \exp\left\{ -\int_{\mathbb{R}^+ \times \mathbb{X}} [1 - e^{-s\, f(y)}] \nu(ds, dy) \right\} \tag{1.5}$$

for all measurable functions $f : \mathbb{X} \to [0, +\infty)$.
We conclude by listing some important classes of CRMs that will be treated in this work: the gamma, the beta and the generalized gamma CRM. We refer to Lijoi & Prünster (2010) for an exhaustive account. A random measure $\tilde{\mu} \sim \text{Ga}(b, \alpha)$ is a gamma CRM of rate parameter $b > 0$ and base measure $\alpha \in \mathcal{M}(\mathbb{X})$ if its Lévy intensity is

$$\nu(ds, dy) = \frac{e^{-sb}}{s} \mathbb{1}_{(0,+\infty)}(s)\, ds\, \alpha(dy). \tag{1.6}$$

With a slight abuse of notation we will sometimes refer to α as the total mass of the base measure, so that $\alpha(dy) = \alpha P_0(dy)$, where $\alpha > 0$ and $P_0 \in \mathcal{P}_{\mathbb{X}}$. We hope that the role of α will be clear from the context.
A random measure $\tilde{\mu} \sim \text{Be}(c, \alpha)$ is a beta CRM of concentration parameter $c > 0$ and base measure $\alpha \in \mathcal{M}(\mathbb{X})$ if

$$\nu(ds, dy) = \frac{c\,(1-s)^{c-1}}{s} \mathbb{1}_{(0,1)}(s)\, ds\, \alpha(dy). \tag{1.7}$$

A random measure $\tilde{\mu} \sim \text{GenGamma}(b, \sigma, \alpha)$ is a generalized gamma CRM of rate parameter $b > 0$, parameter $\sigma \in [0, 1)$ and base measure $\alpha \in \mathcal{M}(\mathbb{X})$ if

$$\nu(ds, dy) = s^{-1-\sigma} e^{-bs} ds\, \alpha(dy). \tag{1.8}$$

1.4.2 Uses in Bayesian Nonparametrics

As mentioned in Section 1.3.3, CRMs can be used to define de Finetti priors through normalization. Let $\tilde{\mu} \sim \text{CRM}(\nu)$ such that $0 < \tilde{\mu}(\mathbb{X}) < +\infty$ almost surely. Then one defines a *normalized random measure with independent increments* as

$$\tilde{P}(\cdot) = \frac{\tilde{\mu}(\cdot)}{\tilde{\mu}(\mathbb{X})} \sim \text{NRMI}(\nu).$$

We will often refer to such priors as normalized CRMs. The Bayesian inference associated with these de Finetti priors has been proposed and studied in Regazzini et al. (2003). In particular the authors prove that if (i) the integrability condition (1.4) holds for every measurable $A \subset \mathbb{X}$ and (ii) for any measurable $A \subset \mathbb{X}$ and for any $\epsilon > 0$, $\nu((0, \epsilon] \times A) = +\infty$, then $0 < \tilde{\mu}(\mathbb{X}) < +\infty$ almost surely. The latter condition is usually addressed to as *inifinite activity* of the CRM. In particular, all the CRMs that we will treat within this thesis will fulfill these standard requirements. This includes gamma random measures: when $\tilde{\mu} \sim \text{Ga}(b, \alpha P_0)$, Ferguson (1973) proved that

$$\frac{\tilde{\mu}(\cdot)}{\tilde{\mu}(\mathbb{X})} \sim \text{DP}(\alpha P_0).$$

Other notable priors that are found through normalization include the normalized stable process (Kingman, 1975), the normalized inverse–Gaussian process (Lijoi et al., 2005) and the normalized generalized gamma process (Brix, 1999; Pitman, 2003).

Normalized random measures with independent increments may be used as de Finetti measures in the same spirit of Theorem 2, i.e.

$$X_1, \dots, X_n | \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}; \qquad \tilde{P} \sim \Pi, \tag{1.9}$$

where Π is the law of a random probability measure (James et al., 2006, 2009). On the other hand, one can also specify the law of the de Finetti measure indirectly, by modeling other characterizing quantities, such as the density function $\tilde{f}(\cdot)$. One of the most popular models is the kernel mixture over a probability measure, which in this context is referred to as mixing measure:

$$X_1, \dots, X_n | \tilde{P} \overset{\text{iid}}{\sim} \tilde{f}(t) = \int_{\mathbb{X}} k(t|x)\, d\tilde{P}(x); \qquad \tilde{P} \sim \Pi. \tag{1.10}$$

This model was first introduced for the Dirichlet process in Ferguson (1983) and Lo (1984) and provides continuous estimates of the density, contrarily to (1.9). One downside is that one loses the closed form expression for the posterior distribution and must content with efficient algorithms to sample from it.
The use of CRMs in Bayesian nonparametrics is not confined to normalization. Indeed, one often specifies the de Finetti measure through the survival function, the hazard function or the cumulative hazards. The latter functions are particular appealing in the context of survival analysis and reliability theory, where CRMs are frequently used as main building block to specify the prior. For example, (Doksum, 1974) provided a convenient way to model the survival function $\tilde{S}(\cdot)$ through neutral to the right processes. The author also showed that these may be characterized as

$$\tilde{S}(t) = e^{-\tilde{\mu}[t,+\infty)},$$

for some CRM $\tilde{\mu}$. Moreover, Hjort (1990) proposed a popular model for the cumulative hazards $\tilde{H}(t)$ by using a beta random measure $\tilde{\mu}$,

$$\tilde{H}(t) = \tilde{\mu}(0, t].$$

As for hazard functions, CRMs may be conveniently used to build kernel mixture models,

$$\tilde{h}(t) = \int_{\mathbb{X}} k(t|x)\, d\tilde{\mu}(x). \tag{1.11}$$

This model was initially proposed with a gamma random measure and a specific kernel by Dykstra & Laud (1981), and has been further generalized to generic kernels (Lo & Weng, 1989) and to generic CRMs (James, 2005).

1.4.3 Approximation

In the previous subsection we have listed some popular models that entail making inference on a class $\mathcal{C}$ of random measures. Since these infinite–dimensional objects are often difficult to treat, both analytically and computationally, in many cases one restricts to

an approximating class $\mathcal{C}^\pi$. The restriction is usually justified with density arguments, typically in terms of weak convergence (Ishwaran & James, 2004; Trippa & Favaro, 2012; Argiento et al., 2016), so that for every $\tilde{\mu} \in \mathcal{C}$ there exists an approximating sequence $\tilde{\mu}_n \in \mathcal{C}^\pi$ such that $\tilde{\mu}_n(A)$ converges to $\tilde{\mu}(A)$ for every Borel set A, as $n \to +\infty$. Nonetheless, with a very few exceptions (Ishwaran & James, 2001; Arbel et al., 2019; Campbell et al., 2019), there is no extensive analysis on how to judge the quality of such approximations. Indeed, the convergence result alone does not give further guidance on how to choose the truncation level n in practice, leading to possibly consistent errors. In order to quantify the approximation error one needs to evaluate

$$d(\tilde{\mu}_n(A), \tilde{\mu}(A)),$$

for each n, where d is a discrepancy measure. In Chapter 2 we will thoroughly address this matter, by taking d equal to the Wasserstein distance. In particular, we will apply our results to a posterior sampling scheme for the hazard model in (1.11), establishing truncation levels based on a precise quantification of the approximation errors.

1.5 Partial exchangeability

Exchangeability may be regarded as a notion of homogeneity between a group of observables. In presence of more than one group of observations, one may often be willing to assume homogeneity within each group, though keeping some heterogeneity across different groups. In this context, partial exchangeability arises as an extremely natural generalization of exchangeability, as first introduced in de Finetti (1938).

Definition 5. Let $m \in \mathbb{N}^+$. A sequence of random variables $\{(X_{i,j})_{j\geq 1} : i = 1, \dots, m\}$ is *partially exchangeable* if for every $(n_1, \dots, n_m) \in \mathbb{N}^m$ and every $(\sigma_1, \dots, \sigma_m) \in \prod_{i=1}^m \mathcal{S}_{n_i}$,

$$\{(X_{i,j})_{j\geq 1} : i = 1, \dots, m\} \overset{\mathrm{d}}{=} \{(X_{i,\sigma_i(j)})_{j\geq 1} : i = 1, \dots, m\}.$$

A generalization of de Finetti's Representation Theorem states that for any partially exchangeable sequence $\{(X_{i,j})_{j\geq 1} : i = 1, \dots, m\}$ there exists a probability measure Π_m on $\mathcal{P}_{\mathbb{X}}^m$ such that for every $n_i \in \mathbb{N}$ and every measurable $A_i \subset \mathbb{X}^{n_i}$ for $i = 1, \dots, m$,

$$\mathbb{P}(\cap_{i=1}^m \{(X_{i,1}, \dots, X_{i,n_i}) \in A_i\}) = \int_{\mathcal{P}_{\mathbb{X}}^{(m)}} \prod_{i=1}^m P_i^{(n_i)}(A_i)\, d\Pi_m(P_1, \dots, P_m),$$

where $P^{(n)} = \prod_{i=1}^n P$ is the n–fold product probability on $\mathbb{X}^n$ with marginal distributions equal to $P \in \mathcal{P}_{\mathbb{X}}$. From a statistical perspective, this characterization is usually rewritten in terms of a dependent random probabilities:

$$\begin{aligned} &(X_{i,1}, \dots, X_{i,n_i})_{i=1}^m | (\tilde{P}_1, \dots, \tilde{P}_m) \sim \tilde{P}_1^{(n_1)} \times \cdots \times \tilde{P}_m^{(n_m)}; \\ &(\tilde{P}_1, \dots, \tilde{P}_m) \sim \Pi_m. \end{aligned}$$

It follows that the law of a partially exchangeable sequence is characterized by the law of a random vector of probabilities. In particular, we point out that when the random probabilities are equal almost surely, we recover the exchangeability assumption. The next section summarizes some popular methods for building dependent random probabilities.

1.6 Dependent random probabilities

Though the first contribution dates to Cifarelli & Regazzini (1978), most proposals for laws of dependent random probabilities have followed the two seminal papers of MacEachern (1999, 2000). Recent years have seen an outburst of models which deal with this inferential problem in creative ways, providing for both analytical and computational tools. In this section we will try to give a brief overview of the subject that by no means aims at being complete. In particular, we focus on models for finite–dimensional vectors of random probabilities, which cover inference for a finite number of groups of observations, i.e. in presence of categorical covariates. This leaves out significant work on spatial or time covariates, which offer an extremely interesting line of research but are not particularly related to the topics of this work. We refer to Quintana et al. (2020) for a recent review.

In Section 1.4.2 we have pointed out how the law of a random probability may be specified directly on $\tilde{P}$ or by resorting to characterizing quantities, such as density, survival function, cumulative hazards and hazard function. Consequently, when defining dependent random probabilities one can chose any of the previous frameworks. When dealing with priors on $\tilde{P}$ or on mixture densities as in (1.10), the starting point is usually the Dirichlet process: as we shall see, most dependent structures were first defined for the Dirichlet process and then extended to other priors. In Section 1.3.3 we have seen how the Dirichlet process may be expressed through the stick–breaking representation and as a normalized CRM, which brings to a breakdown between models for dependent Dirichlet processes. Indeed, one can identify dependence models that rely on the stick–breaking construction, placing the dependence between weights or atoms, and methods that rely on the representation as normalized random measure, defining the dependence at the level of the underlying CRMs. Since the two representations are equivalent, one can sometimes express the same dependence structure in both ways, leveraging the advantages of each representation: the computational advantages of the first and the analytic properties of the second. Before moving to the main examples, we further mention that the models for dependent CRMs may be also used in contexts other than normalization, playing a prominent role in Bayesian nonparametric models for dependent survival functions, hazards and cumulative hazards.

1.6.1 Hierarchical processes

The hierarchical process is arguably one of the most notable tools to introduce dependence between random probability measures. The underlying idea is to assume that the random measures are exchangeable with a de Finetti measure that has full support on some specific class of priors on the space of probabilities $\mathcal{P}_{\mathbb{X}}$, such as the Dirichlet processes. Let $\alpha, \alpha^* > 0$ and let $P^* \in \mathcal{P}_{\mathbb{X}}$. $(\tilde{P}_1, \dots, \tilde{P}_m) \sim \text{HDP}(\alpha, \alpha^*, P^*)$ is a Hierarchical Dirichlet process if

$$\begin{aligned}\tilde{P}_1, \dots, \tilde{P}_m | \tilde{P} &\sim \text{DP}(\alpha, \tilde{P});\\ \tilde{P} &\sim \text{DP}(\alpha^*, P^*).\end{aligned}$$

This dependence structure was first introduced in Teh et al. (2006) starting from the stick–breaking characterization and subsequently expressed in terms of normalization in Camerlenghi et al. (2019b). In particular, the authors extended the definition of hierarchical process to CRMs. As a side effect, they were also able to study many distributional properties of the hierarchical Pitman–Yor process, which we define in a natural way as follows. Let $\alpha, \alpha^* \in [0, 1)$, $\theta, \theta^* > 0$ and $P^* \in \mathcal{P}_{\mathbb{X}}$. Then $(\tilde{P}_1, \dots, \tilde{P}_m) \sim \text{HPY}(\alpha, \theta, \alpha^*, \theta^*, P^*)$ if

$$\begin{aligned}\tilde{P}_1, \dots, \tilde{P}_m | \tilde{P} &\sim \text{PY}(\alpha, \theta, \tilde{P});\\ \tilde{P} &\sim \text{PY}(\alpha^*, \theta^*, P^*).\end{aligned}$$

Hierarchical CRMs have also been used in the context of survival analysis in Camerlenghi et al. (2020).

1.6.2 Nested processes

Another popular way to model exchangeable random probabilities is through the nested process, where the de Finetti measure is distributed as a Dirichlet process whose base space is the space of probabilities $\mathcal{P}_{\mathbb{X}}$. Let $\alpha, \alpha^* > 0$ and let $P^* \in \mathcal{P}_{\mathbb{X}}$. $(\tilde{P}_1, \dots, \tilde{P}_m) \sim \text{NDP}(\alpha, \alpha^*, P^*)$ is a nested Dirichlet process if

$$\begin{aligned}\tilde{P}_1, \dots, \tilde{P}_m | \tilde{\Pi} &\sim \tilde{\Pi};\\ \tilde{\Pi} &\sim \text{DP}(\alpha, \text{DP}(\alpha^*, P^*)).\end{aligned}$$

This model was first defined in Rodrguez et al. (2008) for the stick–breaking representation and then extended to normalized CRMs in Camerlenghi et al. (2019a).

1.6.3 Additive processes

Additive dependence structures model the nonparametric priors as convex sums of a common component and an idiosyncratic one. Let $\alpha > 0$,$P_0 \in \mathcal{P}_{\mathbb{X}}$ and $\lambda_1, \dots, \lambda_m \in [0, 1]$. Then $(\tilde{P}_1, \dots, \tilde{P}_m)$ is an additive Dirichlet process if

$$\begin{aligned}\tilde{P}_i &= \lambda_i \tilde{Q}_0 + (1 - \lambda_i)\tilde{Q}_i;\\ \tilde{Q}_0, \tilde{Q}_1, \dots, \tilde{Q}_n &\overset{\text{iid}}{\sim} \text{DP}(\alpha, P_0).\end{aligned}$$

This model was first proposed by Müller et al. (2004) for the stick–breaking representation and then extended in Lijoi et al. (2014) to normalized random measures. In particular, the authors define additive CRMs by introducing the dependence at the level of the Poisson processes, following an additive dependence structure proposed in Griffiths & Milne (1978). For this reason, additive CRMs are often referred to as GM-dependent. These were then used in Lijoi & Nipoti (2014) to model dependent hazards.

1.6.4 Compound random measures

Many common procedures introduce dependence between CRMs by relying on random vectors $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \dots, \tilde{\mu}_m)$ with jointly independent increments, i.e. such that given a finite collection of disjoint bounded Borel sets $\{A_1, \cdots, A_n\}$, the random elements $\{\tilde{\boldsymbol{\mu}}(A_1), \dots, \tilde{\boldsymbol{\mu}}(A_n)\}$ are independent. When this condition holds, we refer to $\tilde{\boldsymbol{\mu}}$ as a completely random vector (CRV). In particular this entails that the marginal random measures of a CRV are completely random, though we point out that the converse is not necessarily true: a random vector of measures whose marginals are CRMs is not necessarily a CRV. The independence of the increments guarantees that if $\tilde{\boldsymbol{\mu}}$ has no fixed atoms, there exists a Poisson random measure $\mathcal{N}$ on $\mathbb{R}_+^m \times \mathbb{X}$ such that for every $A \in \mathcal{X}$,

$$\tilde{\boldsymbol{\mu}}(A) \stackrel{\mathrm{d}}{=} \int_{\mathbb{R}_+^m \times A} \boldsymbol{s}\, \mathcal{N}(ds_1, \dots, ds_m, dx), \tag{1.12}$$

where $\boldsymbol{s} = (s_1, \dots, s_m)$. Thus, in order to specify a law for dependent CRMs it suffices to provide a multivariate Lévy intensity $\nu(ds_1, \dots, ds_m, dx) = \mathbb{E}(\mathcal{N}(ds_1, \dots, ds_m, dx))$. Among these methods, compound random measures stand out as a particularly tractable approach that was first introduced in Griffin & Leisen (2017), though a special case may be already be found in Leisen et al. (2013). The multivariate Lévy intensity is modeled as

$$\nu(ds_1, \dots, ds_m, dx) = \alpha P_0(dx) \int_{\mathbb{R}^+} \frac{1}{u^2}\, h\left(\frac{s_1}{u}, \dots, \frac{s_m}{u}\right) \nu^*(du)\, ds_1 \dots ds_m,$$

where $\alpha > 0$, $P_0 \in \mathcal{P}_{\mathbb{X}}$, $h : \mathbb{R}^m \to \mathbb{R}^+$ is a density function and $\alpha\nu^*(du)P_0(dx)$ is a Lévy intensity on $\mathbb{R}^+ \times \mathbb{R}$. Adequate choices for ν^* and h bring to different dependence structures and marginal CRMs, including gamma, generalized gamma, beta and stable random measures. Moreover the authors also provided a series representation that highlights how compound random measures have dependent jumps and share the same atoms. These have been used both in the context of density estimation through normalization (Griffin & Leisen, 2018) and in the context of survival analysis (Riva Palacio & Leisen, 2018).

1.6.5 Lévy copulae

Lévy copulae provide another effective way to model multivariate Lévy intensities. As copulae can be seen as a way to separate the marginal components of a bivariate distribution from its dependence structure, so do their generalization to Lévy intensities,

conceived in Tankov (2003) and Cont & Tankov (2004) to build dependent Lévy processes. For simplicity, we focus on 2–dimensional copulae, though the definition and the main results may be extended to a generic $m \geq 2$. A Lévy copula is a function $c : [0, +\infty]^2 \to [0, +\infty]$ such that

1. $\forall u \in [0, +\infty]$, $c(u, 0) = c(0, u) = 0$;
2. $\forall u \in [0, +\infty]$, $c(+\infty, u) = c(u, +\infty) = u$;
3. $\forall\ 0 \leq u_1 \leq u_2,\ 0 \leq v_1 \leq v_2$,
   $c(u_2, v_2) - c(u_2, v_1) - c(u_1, v_2) + c(u_1, v_1) \geq 0$.

Given a Lévy intensity on $\mathbb{R}_+^2 \times \mathbb{X}$, $\nu(ds_1, ds_2, A) = \int_{\mathbb{X}} \mathbb{1}_A(x)\nu(ds_1, ds_2, dx)$ is a Lévy intensity on $\mathbb{R}_+^2$ for any measurable $A \subset \mathbb{X}$. We indicate its marginal tail integrals by $U_{1,A}(t) = \nu([t, +\infty) \times \mathbb{R}_+ \times A)$ and $U_{2,A}(t) = \nu(\mathbb{R}_+ \times [t, +\infty) \times A)$. An analogue of Sklar's theorem states that there exists a Lévy copula $c : [0, +\infty]^2 \to [0, +\infty]$ such that

$$\nu((t_1, +\infty) \times (t_2, +\infty) \times A) = c(U_{1,A}(t_1), U_{2,A}(t_2)).$$

When the Lévy copula and the tail integrals are sufficiently smooth, $\nu(ds_1, ds_2, A)$ is recovered by

$$\nu(ds_1, ds_2, A) = \frac{\partial^2}{\partial u_1 \partial u_2} c(u_1, u_2)\big|_{U_{1,A}(s_1), U_{2,A}(s_2)}\, \nu_1(ds_1, A)\, \nu_2(ds_2, A). \tag{1.13}$$

Lévy copulae are a useful instrument to build multivariate Lévy intensities with a close look at their dependence structure. They have been used in the context of survival analysis in (Epifani & Lijoi, 2010) and through normalization in Leisen & Lijoi (2011).

1.7 Measuring dependence

In the previous section we have underlined how many popular models for partially exchangeable observables are based on dependent random measures. The amount of dependence plays a fundamental role in the learning mechanism, allowing for different degrees of borrowing of information across groups. One can identify two extreme prior assumptions: on one hand when the random measures are independent, so are the groups of observables, so that there will be no sharing of information between them; on the other hand, when the random measures are almost surely equal, the observations are exchangeable and there is maximal borrowing of strength across groups. A precise elicitation of the prior dependence structure is crucial for these models, since it has a great impact on the learning mechanism and the posterior inference. Interestingly, a principled approach to dependence is still missing is this context and one usually considers measures of set–wise linear correlation, which in presence of two groups amounts to $\text{cor}(\tilde{\mu}_1(A), \tilde{\mu}_2(A))$, for any Borel set A. In Chapter 3 we propose a way to go beyond linear correlation, by introducing a non-model specific quantification of dependence in terms of closeness to exchangeability. To this end, we recall that exchangeability is recovered whenever the

random measures are equal almost surely, i.e. $\tilde{\mu}_1 = \tilde{\mu}_2$ a.s. We will refer to this vector as comonotonic and denote it as $\tilde{\boldsymbol{\mu}}^{\text{co}}$. We then propose to measure the dependence of a vector $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \tilde{\mu}_2)$ as a distance $d(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}^{\text{co}})$ from the comonotonic vector. We will build d on the Wasserstein distance because it is ideally suited for measuring discrepancy between random quantities and, more specifically, it allows for an informative comparison between distributions with different support. This will then be used for quantifying the dependence structure in noteworthy Bayesian nonparametric models, including additive structures (1.6.3), compound random measures (1.6.4) and Lévy copulae (1.6.5).

Chapter 2

Approximation of Bayesian models for time-to-event data

Random measures are the key ingredient for effective nonparametric Bayesian modeling of time-to-event data. This chapter focuses on priors for the hazard rate function, a popular choice being the kernel mixture with respect to a gamma random measure. Sampling schemes are usually based on approximations of the underlying random measure, both *a priori* and conditionally on the data. Our main goal is the quantification of approximation errors through the Wasserstein distance. Though easy to simulate, the Wasserstein distance is generally difficult to evaluate, making tractable and informative bounds essential. Here we accomplish this task on the wider class of completely random measures, yielding a measure of discrepancy between many noteworthy random measures, including the gamma, generalized gamma and beta families. By specializing these results to gamma kernel mixtures, we achieve upper and lower bounds for the Wasserstein distance between hazard rates, cumulative hazard rates and survival functions.

2.1 Introduction

One of the most attractive features of the Bayesian nonparametric approach to statistical inference is the modeling flexibility implied by priors with large support. There are several classes of priors where this property is complemented by analytical tractability, thus contributing to making Bayesian nonparametrics very popular in several applied areas. See Hjort et al. (2010) and Ghosal & van der Vaart (2017) for broad overviews. In this framework, survival analysis stands out as one of the most lively fields of application. A prominent model for exchangeable time-to-event data is the extended gamma process for hazard rates (Dykstra & Laud, 1981), which allows for continuous observables and has been further generalized to kernel mixtures in Lo & Weng (1989) and James (2005). These works paved the way for another active line of research that defines priors for the hazard rates by relaxing the dependence structure between the observables, going beyond the exchangeability assumption. For example, Pennell & Dunson (2006), De Iorio et al. (2009) and Hanson et al. (2012) model subject specific hazards based

on continuous covariates; Lijoi & Nipoti (2014) and Zhou et al. (2015) define priors for cluster specific hazards, while Nipoti et al. (2018) account for both individual and cluster covariates simultaneously. In this chapter we rather focus on priors for the hazard rates of exchangeable time-to-event data.

An important feature shared by most classes of nonparametric priors is their definition in terms of random measures and transformations thereof. While there is a wealth of theoretical results that have eased their actual implementation in practice, sampling schemes are typically based on approximations of the underlying random measures. Nonetheless, with a very few exceptions (Ishwaran & James, 2001; Arbel et al., 2019; Campbell et al., 2019), there is no extensive analysis on how to judge the quality of such approximations. Consider the common situation where one is interested in making inference or sampling from a wide class of random measures $\mathcal{C}$, but can only treat a subclass $\mathcal{C}^{\pi}$ because of good analytical or computational properties. The restriction to $\mathcal{C}^{\pi}$ is usually argued through density statements, typically in terms of weak convergence of random measures. In many cases this reduces to the weak convergence of one–dimensional distributions, i.e. for every $\tilde{\mu} \in \mathcal{C}$ there exists an approximating sequence $\{\tilde{\mu}_n\}_{n\geq 1}$ in $\mathcal{C}^{\pi}$ such that $\tilde{\mu}_n(A)$ converges weakly to $\tilde{\mu}(A)$ for every Borel set A. This leaves out the possibility of establishing the rate of convergence and, more importantly, provides no guidance on the choice of the approximation $\tilde{\mu}_{\bar{n}} \in \{\tilde{\mu}_n\}_{n\geq 1}$ to use in practical implementations. Spurred by these considerations, the goal we pursue is to quantify the approximation errors by evaluating the Wasserstein distance between $\tilde{\mu}_n(A)$ and $\tilde{\mu}(A)$. Since convergence in Wasserstein distance implies weak convergence, this has the additional advantage of strengthening most results known in the literature.

The Wasserstein distance was first defined by Gini (1914) as a *simple measure of discrepancy* between random variables. During the 20^{th} century it has been redefined and studied in many other disciplines, such as transportation theory, partial differential equations, ergodic theory and optimization. Nowadays, depending on the field of study, it is known with different names, such as Gini distance, coupling distance, Monge-Kantorovich distance, Earth Moving distance and Mallows distance; see Villani (2008), Rachev (1985) and Cifarelli & Regazzini (2017) for reviews. Indeed, one can find it scattered across the statistics literature (Mallows, 1972; Dudley, 1976; Bickel & Freedman, 1981; Chen, 1995), though only in recent years it has achieved major success, especially in probability and machine learning. For a detailed review on the uses of the Wasserstein distance in statistics see Panaretos & Zemel (2019). As for the Bayesian literature, the Wasserstein distance was first used in Nguyen (2013) and has been mainly used to evaluate approximations of the posterior distribution and to deal with consistency (Nguyen, 2013; Srivastava et al., 2015; Cifarelli et al., 2016; Gao & van der Vaart, 2016; Donnet et al., 2018; Heinrich & Kahn, 2018). These works deal with the convergence of the (random) Wasserstein distance between the attained values of random probability measures. In a similar vein, though without a specific statistical motivation, Mijoule et al. (2016) examine the Wasserstein convergence rate of the empirical distribution to the prior, namely the de Finetti measure, for an exchangeable sequence of $\{0,1\}$–valued random variables. Our approach goes in a different direction: we are interested in a

distance between the laws of random measures rather than a random distance between measures.
The Wasserstein distance may be effectively approximated through simulation (Sriperumbudur et al., 2012; Cuturi, 2013) but it is difficult to evaluate it analytically and, hence, tractable bounds are needed for concrete applications. We achieve them in two steps. First, we determine bounds for the Wasserstein distance between so–called *completely random measures*, since they act as building blocks of most popular nonparametric priors. This is carried out by relying on results in Mariucci & Reiß (2018) on Lévy processes. The techniques we develop in this first part measure the discrepancy between the laws of many noteworthy random measures, including the gamma, generalized gamma and beta families. Secondly, we move on to using these bounds in order to quantify the divergence between hazard rate mixture models that are used to model time-to-event data. These are then applied to evaluate the approximation error in a posterior sampling scheme for the hazards, in multiplicative intensity models, that relies on an algorithm for extended gamma processes (Al Masry et al., 2017).

The outline of the chapter is as follows. After providing some basic notions and results on the Wasserstein distance and on completely random measures in Section 2.2, we determine upper and lower bounds for the Wasserstein distance between one–dimensional distributions associated to completely random measures in Section 2.3. This is specialized to the case of gamma and beta completely random measures in Section 2.3.2. These results are the starting point for carrying out an in–depth analysis of hazard rate mixture models driven by completely random measures. In Section 2.4 we obtain a quantification of the discrepancy between two hazard rate mixtures and for the associated random survival functions. Examples related to its specification with mixing gamma random measures may be found in Section 2.4.3. Finally, in Section 2.5 we apply these results to evaluate the approximation error of a sampling scheme for the posterior hazards, conditional on the data. Proofs of the main results are deferred to Section 2.6.

2.2 Convergence of completely random measures

In this first section we recall notions about the convergence of completely random measures in terms of the Wasserstein distance.
Let $\mathbb{X}$ be a Polish space with distance $d_{\mathbb{X}}$ and Borel σ-algebra $\mathcal{X}$. The space $M_{\mathbb{X}}$ of boundedly finite measures on $\mathbb{X}$ endowed with the weak$^\sharp$ topology is a Polish space as well; see Daley & Vere-Jones (2002). We denote by $\mathscr{M}_{\mathbb{X}}$ the corresponding Borel σ-algebra and consider completely random measures (CRMs $\tilde{\mu} : (\Omega, \Sigma, \mathbb{P}) \to (M_{\mathbb{X}}, \mathscr{M}_{\mathbb{X}})$, as defined in Section 1.4. Motivated by Bayesian nonparametric modeling, we focus on CRMs without fixed atoms such that the corresponding Lévy intensity ν satisfies

$$\int_A \int_{\mathbb{R}^+} (\epsilon \wedge s)\, \nu(ds, dy) < +\infty \tag{2.1}$$

for every A in $\mathcal{X}$ and such that infinitely activity holds, i.e. for all A in $\mathcal{X}$ and $\epsilon > 0$, $\nu((0, \epsilon] \times A) = +\infty$. As mentioned in Section 1.4.2, this is a standard requirement for most applications in Bayesian nonparametrics.
When dealing with convergence of random measures we think of random measures in terms of probability distributions on $M_{\mathbb{X}}$. Results in strong convergence are often too hard to establish, so that one usually deals with weak convergence (of distributions) of random measures, $\mathcal{L}(\tilde{\mu}_n) \Rightarrow \mathcal{L}(\tilde{\mu})$, where $\mathcal{L}(X)$ denotes the probability distribution of a random element X, which can be either finite– or infinite–dimensional. A remarkable result establishes that this is equivalent to the weak convergence of all finite–dimensional distributions $\mathcal{L}(\tilde{\mu}_n(A_1), \cdots, \tilde{\mu}_n(A_d)) \Rightarrow \mathcal{L}(\tilde{\mu}(A_1), \cdots, \tilde{\mu}(A_d))$, for $A_1, \ldots, A_d \in \mathcal{X}$ stochastic continuity sets for $\tilde{\mu}$; see Theorem 11.1.VII in Daley & Vere-Jones (2007). Moreover, when dealing with CRMs, the weak convergence of finite–dimensional distributions is equivalent to the weak convergence of one–dimensional distributions. Thus, if d denotes a metric on the probability distributions on $\mathbb{R}$ whose convergence is stronger than the weak convergence, one has that if

$$d(\mathcal{L}(\tilde{\mu}_n(A)), \mathcal{L}(\tilde{\mu}(A))) \to 0 \tag{2.2}$$

for every $A \in \mathcal{X}$, then $\tilde{\mu}_n$ converges weakly to $\tilde{\mu}$. In the sequel, we will choose d as the Wasserstein distance of order 1 with respect to the Euclidean norm on $\mathbb{R}$, as defined in Section A in the appendix. We denote such distance by $\mathcal{W}$ and refer to Villani (2008) for a complete reference. In view of the previous discussion on weak convergence, a major goal that we pursue is evaluating or bounding $\mathcal{W}(\tilde{\mu}_1(A), \tilde{\mu}_2(A))$, for A in $\mathcal{X}$. Before detailing the results and general techniques we rely on in order to achieve them in the next few sections, we recall that for any pair of random variables (X, Y),

$$|\mathbb{E}(X) - \mathbb{E}(Y)| \leq \mathcal{W}(X, Y) \leq \mathbb{E}(|X|) + \mathbb{E}(|Y|), \tag{2.3}$$

as in Proposition 52 in the appendix. Thus the Wasserstein distance is finite when the random variables have finite mean. We will therefore focus our attention on CRMs whose total mass has finite mean. By Campbell's theorem this boils down to

$$\mathbb{E}(\tilde{\mu}(\mathbb{R})) = \mathbb{E}\left(\int_{\mathbb{R}^+ \times \mathbb{X}} s \, \mathcal{N}(ds, dy) \right) = \int_{\mathbb{R}^+ \times \mathbb{X}} s \, \nu(ds, dy) < +\infty. \tag{2.4}$$

2.3 Wasserstein bounds for completely random measures

2.3.1 General result

There are situations where one is only interested in a numerical value for the Wasserstein distance: in such a case there are efficient ways to simulate it (Sriperumbudur et al., 2012; Cuturi, 2013). On the other hand, one may be interested in understanding how the distance is affected by the choice of distributions, by the parameters or by meaningful functionals, such as moments. This raises the need for an analytical evaluation of the Wasserstein distance, which in general is not an easy task. The most common practice is

thus to develop meaningful bounds and to analyze how these are affected by the choices above. In this section we will express a bound for the Wasserstein distance between the one–dimensional distributions of CRMs in terms of their corresponding Lévy intensities. The proof is based on a compound Poisson approximation of CRMs.

Theorem 5. *Let $\tilde{\mu}_1$ and $\tilde{\mu}_2$ be infinitely active CRMs with finite total mean. Then for every $A \in \mathcal{X}$*

$$g_\ell(A) \leq W(\tilde{\mu}_1(A), \tilde{\mu}_2(A)) \leq g_u(A),$$

where

$$g_\ell(A) = |\mathbb{E}(\tilde{\mu}_1(A)) - \mathbb{E}(\tilde{\mu}_2(A))| = \left| \int_{\mathbb{R}^+} s\, \nu_1(ds \times A) - \int_{\mathbb{R}^+} s\, \nu_2(ds \times A) \right|;$$

$$g_u(A) = \int_0^{+\infty} |\nu_1((u, +\infty) \times A) - \nu_2((u, +\infty) \times A)|\, du.$$

We observe that $g_u(A)$ has a compelling form with respect to the upper bound in (2.3), since it equals zero if $\tilde{\mu}_1 \overset{\mathrm{d}}{=} \tilde{\mu}_2$. We stress that this bound holds for all CRMs and may be evaluated through numerical integration. Nonetheless, when specializing to certain classes of CRMs, we manage to bound $g_u(A)$ from above with an expression that can be evaluated exactly, as we do in Section 2.3.2. In particular, easily computable upper bounds are available whenever the tails of the Lévy intensities are ordered, as we prove in the next corollary. In such a case not only we have a simple expression for $g_u(A)$, we may also prove that the upper and lower bounds coincide, providing the exact expression of the Wasserstein distance itself.

Corollary 6. *Consider the hypotheses of* Theorem 5 *and let $A \in \mathcal{X}$. If the tails of $\nu_1(ds \times A)$ and $\nu_2(ds \times A)$ are ordered, namely $\nu_i((u, +\infty) \times A) \leq \nu_j((u, +\infty) \times A)$ for all $u \in \mathbb{R}^+$ and $i \neq j$ in $\{1, 2\}$, then*

$$\mathcal{W}(\tilde{\mu}_1(A), \tilde{\mu}_2(A)) = \left| \int_{\mathbb{R}^+} s\, \nu_1(ds \times A) - \int_{\mathbb{R}^+} s\, \nu_2(ds \times A) \right|.$$

Remark 1. The condition of Corollary 6 holds whenever there exists a dominating measure η on $\mathbb{R}^+$ such that the Radon–Nikodym derivatives of $\nu_1(ds \times A)$ and $\nu_2(ds \times A)$ are ordered, i.e. $\nu_{i,A}(s) \leq \nu_{j,A}(s)$ for all $s \in \mathbb{R}^+$ and $i \neq j$ in $\{1, 2\}$. This more restrictive condition, which is however much easier to verify, holds true for many examples to be displayed in the sequel.

As underlined in Section 2.2, the convergence in the Wasserstein distance between $\tilde{\mu}_n(A)$ and $\tilde{\mu}(A)$ for every $A \in \mathcal{X}$ is sufficient to guarantee the weak convergence of the CRMs and provide convergence rates. This motivates our main interest in set–wise results as those in Corollary 6. However, one can also define a uniform distance between laws of random measures with finite total mean, by considering

$$d_{\mathcal{W}}(\tilde{\mu}_1, \tilde{\mu}_2) = \sup_{A \in \mathcal{X}} \mathcal{W}(\tilde{\mu}_1(A), \tilde{\mu}_2(A)). \tag{2.5}$$

Corollary 6 can be used to find the exact expression of such distance. We focus on *homogeneous* CRMs, i.e. such that their Lévy intensity is a product measure $\nu(ds, dx) = \rho(s)\, ds\, \alpha(dx)$. It will be next shown that $d_{\mathcal{W}}$ admits a very intuitive representation, being proportional to the total variation distance between the base measures

$$\mathrm{TV}(\alpha_1, \alpha_2) = \sup_{A \in \mathcal{X}} |\alpha_1(A) - \alpha_2(A)|.$$

Corollary 7. *Let $\tilde{\mu}_i$ be infinitely active homogeneous CRMs with finite total mean such that the Lévy intensities $\nu_i(ds, dx) = \rho(s)\, ds\, \alpha_i(dx)$, for $i = 1, 2$. Then,*

$$d_{\mathcal{W}}(\tilde{\mu}_1, \tilde{\mu}_2) = \mathrm{TV}(\alpha_1, \alpha_2) \int_{\mathbb{R}^+} s\, \rho(s)\, ds.$$

2.3.2 Examples

When the conditions of Corollary 6 do not hold, we may often find upper bounds of $g_u(A)$ which may be evaluated exactly for specific examples of CRMs. In the next proposition we consider a gamma CRMs, as defined in (1.6) in Section 1.4. A random measure $\tilde{\mu} \sim \mathrm{Ga}(b, \alpha)$ is easily shown to be infinitely active and, if α is a finite measure on $\mathbb{X}$, it has finite total mean.

Proposition 8. *Let $\tilde{\mu}_i \sim Ga(b_i, \alpha_i)$, where $0 < b_1 < b_2$ and α_i is a finite measure on $\mathbb{X}$ for $i = 1, 2$. Then,*

$$g_\ell(\boldsymbol{b}, \boldsymbol{\alpha}, A) \leq \mathcal{W}(\tilde{\mu}_1(A), \tilde{\mu}_2(A)) \leq g_u(\boldsymbol{b}, \boldsymbol{\alpha}, A),$$

where

$$g_\ell(\boldsymbol{b}, \boldsymbol{\alpha}, t) = \left| \frac{\alpha_1(A)}{b_1} - \frac{\alpha_2(A)}{b_2} \right|;$$

$$g_u(\boldsymbol{b}, \boldsymbol{\alpha}, t) = \frac{\alpha_1(A)}{b_1} - \frac{\alpha_2(A)}{b_2} + \mathbb{1}_{(0,+\infty)}(\alpha_2(A) - \alpha_1(A))\, 2\, \frac{\alpha_2(A) - \alpha_1(A)}{b_2 - b_1} \log \frac{b_2}{b_1};$$

and we have used the vector notations $\boldsymbol{b} = (b_1, b_2)$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$.

This result extends the ones in Gairing et al. (2015), who develop upper bounds for similar integrals of gamma Lévy measures in a more restrictive framework as they do not allow for both base measures and the scale parameter to differ between the two specifications. The bounds of Proposition 8 are informative in the sense that, the closer the parameters of the two CRMs, the smaller the bound of the Wasserstein distance. Moreover, when the base measures are equal on A, the upper and lower bounds coincide, providing the exact expression for the Wasserstein distance, in accordance with Corollary 6. The same holds true if $b_1 = b_2$, since

$$\lim_{b_2 \to b_1^+} \frac{1}{b_2 - b_1} \log \frac{b_2}{b_1} = \frac{1}{b_1}.$$

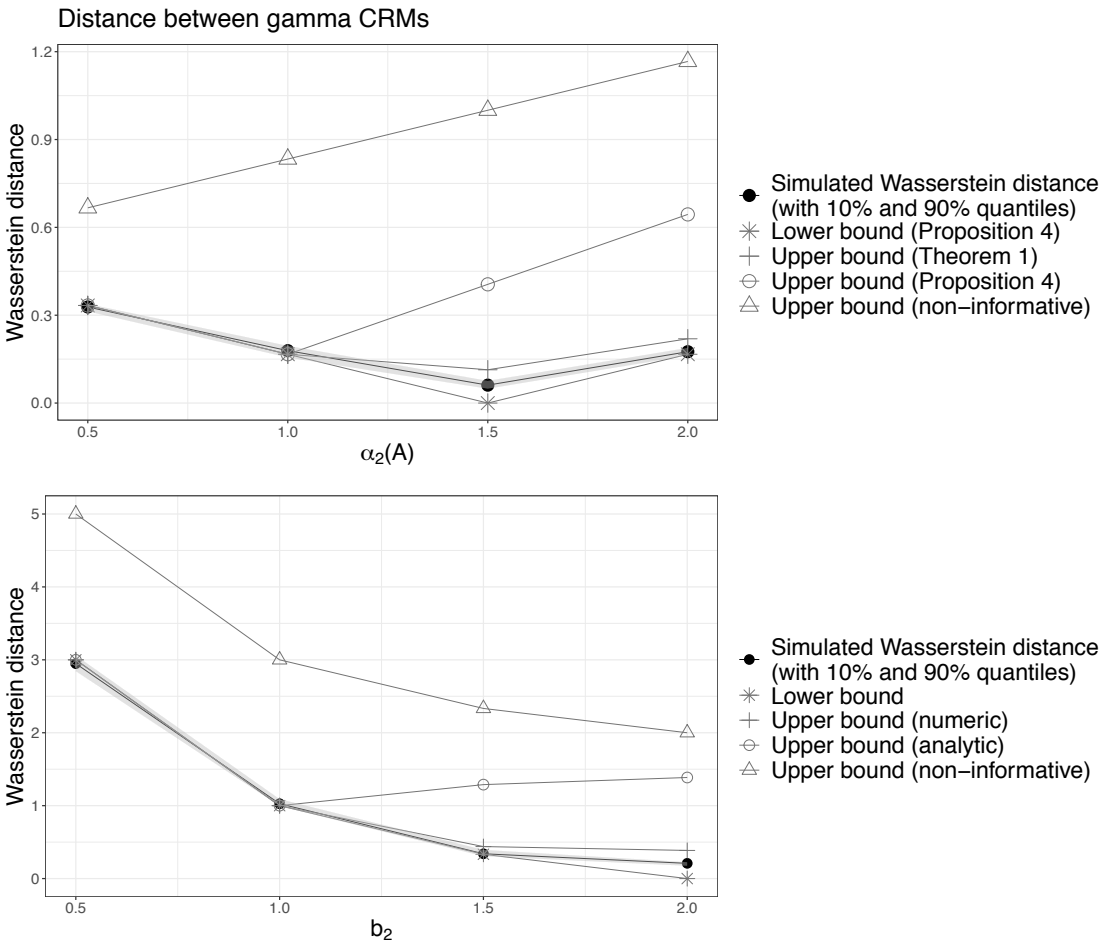

Figure 2.1: Wasserstein distance $\mathcal{W}(\tilde{\mu}_1(A), \tilde{\mu}_2(A))$ between gamma CRMs and relative upper and lower bounds. In the upper panel $\alpha_1(A) = 1, b_1 = 2, b_2 = 3$ are fixed, whereas $\alpha_2(A)$ ranges from 0.5 to 2. In the lower one $\alpha_1(A) = 1, b_1 = 1, \alpha_2(A) = 2$ are fixed, whereas b_2 ranges from 0.5 to 2. In both plots the *Simulated Wasserstein distance* is based on 10 samples of 1000 observations using the Python Optimal Transport (POT) package (Flamary & Courty, 2017).

In Figure 2.1 we compare the simulated Wasserstein distance between two gamma CRMs with the upper bound in Theorem 5, which can be evaluated numerically, the ones in Proposition 8, which can be evaluated exactly, and the upper bound in (2.3), which is non–informative. For a wide range of parameters the bounds of Theorem 5 and Proposition 8 coincide with the Wasserstein distance. In contrast, when the Lévy intensities are not ordered, the upper and lower bounds do not coincide. The upper bound of Proposition 8 is tight whenever at least one of the two parameters is close to the corresponding parameter of the other CRM (i.e. $\alpha_1(A) \approx \alpha_2(A)$ or $b_1 \approx b_2$), whereas the upper bound of Theorem 5 is tight on the whole range of parameters. Moreover, they are both more informative than the upper bound in (2.3). The lower bound, on the other hand, is al-

ways tight and becomes non–informative when the two CRMs have different parameters but equal ratios $(\alpha_i(A)/b_i)$, i.e. when they have equal mean.

A different situation occurs with beta CRMs, defined in (1.7) in Section 1.4, where the Lévy densities corresponding to different concentration parameters and same base measure are not ordered.

Proposition 9. *Let $\tilde{\mu}_i \sim Be(c_i, \alpha_i)$, where $0 < c_1 \leq c_2$ and α_i is a finite measure on $\mathbb{X}$ for $i = 1, 2$. Then,*

1. *If $c_1 = c_2 = c$, then $\mathcal{W}(\tilde{\mu}_1(A), \tilde{\mu}_2(A)) = |\alpha_1(A) - \alpha_2(A)|$.*
2. *If $\alpha_1 = \alpha_2 = \alpha$, then $\mathcal{W}(\tilde{\mu}_1(A), \tilde{\mu}_2(A))) \leq 2\,\alpha(A)\,\log\left(\frac{c_2}{c_1}\right)$*

We conclude this section with an immediate application of Corollary 7 on the distance $d_{\mathcal{W}}$ between the laws of generalized gamma CRMs.

Example 1. Consider $\tilde{\mu}_i \sim \text{GenGamma}(b, \sigma, \alpha_i)$ generalized gamma CRMs, as defined in (1.8) in Section 1.4, for $i = 1, 2$. Then Corollary 7 ensures that $d_{\mathcal{W}}(\tilde{\mu}_1, \tilde{\mu}_2) = \Gamma(1 - \sigma)b^{\sigma-1}\,\text{TV}(\alpha_1, \alpha_2)$. When $\sigma = 0$ we recover the distance between two gamma CRMs with same rate parameter.

2.4 Hazard rate mixtures

Applications in survival analysis and reliability involve *time-to-event* data and have spurred important developments in Bayesian nonparametric modeling. Stimulating and exhaustive overviews of popular models in the area can be found in Müller et al. (2015) and in Ghosal & van der Vaart (2017). If $T_1, \ldots, T_n$ are from an exchangeable sequence of time-to-event data, i.e.

$$T_i \,|\, \tilde{P} \overset{\text{iid}}{\sim} \tilde{P} \quad (i = 1, \ldots, n), \qquad \tilde{P} \sim \Pi, \tag{2.6}$$

the choice of Π follows from specifying a prior on the survival function $t \mapsto \tilde{S}(t) = \tilde{P}((t, \infty))$. This may be done directly by resorting, e.g., to neutral to the right random probability measure (Doksum, 1974), or by setting a prior on the corresponding cumulative hazard function by means of, e.g., the Beta process (Hjort, 1990), or on the hazard rate function if one can assume that $\tilde{S}$ is almost surely continuous, in which case a convenient option is a kernel mixture model (Dykstra & Laud, 1981). For all these model specifications, one can also take into account the presence of censored observations. The most common mechanism is right–censoring, which associates to each T_i a censoring time C_i. In this case, the actual observations are the pairs (X_i, Δ_i), where $X_i = T_i \wedge C_i$ and $\Delta_i = \mathbb{1}_{(0,C_i]}(T_i)$ identifies exact observations whenever it equals 1. Here we focus on priors for the hazard rates, i.e. the instantaneous risk of failure, that are induced by kernel mixtures over a gamma CRM. The model, originally proposed as a prior for increasing hazard rates in Dykstra & Laud (1981), is conjugate with respect to right

censored observations and has led to several interesting generalizations. Henceforth, we consider a specification that has been investigated in its full generality by James (2005). Before focusing on our main results, let us first recall some basic definitions that will also allow us to set the notation to be used throughout. If F is a cumulative distribution function on $[0, +\infty)$ and $S = 1 - F$ the corresponding survival function, we assume it is absolutely continuous so that one can define the *hazard rate* $h = F'/(1-F)$ and rewrite, for any $t \geq 0$,

$$S(t) = \exp\{-H(t)\}; \qquad H(t) = \int_0^t h(s)\, ds,$$

where H is the cumulative hazard function. Let $k : \mathbb{R}^+ \times \mathbb{X} \to [0, +\infty)$ be a measurable kernel function. If $\tilde{\mu}$ is a CRM, with corresponding Poisson random measure $\mathcal{N}$, and k is such that

$$\lim_{t\to\infty} \int_0^t \int_{\mathbb{R}^+\times\mathbb{X}} k(u \,|\, y)\, s\, du\, \mathcal{N}(ds, dy) = +\infty, \tag{2.7}$$

a prior for the hazard rates is the probability distribution of the process $\{\tilde{h}(t) \,|\, t \geq 0\}$ such that for any $t \geq 0$,

$$\tilde{h}(t) = \int_{\mathbb{X}} k(t \,|\, y)\, \tilde{\mu}(dy) \overset{\mathrm{d}}{=} \int_{\mathbb{R}^+\times\mathbb{X}} k(t \,|\, y)\, s\, \mathcal{N}(ds, dy). \tag{2.8}$$

Thus, condition (2.7) ensures that the mean cumulative hazards go to $+\infty$ as time increases. We use the techniques developed in the previous sections to obtain bounds on the Wasserstein distance between the marginal hazard rates coming from different kernels and different CRMs. Moreover, we successfully address the same issue when considering the Wasserstein distance between cumulative hazards and survival functions.

2.4.1 Bounds for hazard rates

Consider two random hazard rates $\tilde{h}_1 = \{\tilde{h}_1(t) \,|\, t \geq 0\}$ and $\tilde{h}_2 = \{\tilde{h}_2(t) \,|\, t \geq 0\}$ as in (2.8). From a statistical standpoint, these may be seen as different prior specifications corresponding, e.g., to different mixing CRMs or kernels. Alternatively, $\tilde{h}_2$ may be thought as an approximation of the actual prior $\tilde{h}_1$ and one may be interested to ascertain the quality of such an approximation. The issue is of great interest when we need to sample from $\tilde{h}_1$, or its posterior distribution, while it is much easier to sample from $\tilde{h}_2$: in this case a bound on the error can provide an effective guidance as on how to fix the parameters of the approximating distribution. We first investigate how different CRMs and kernels impact the marginal hazards and use the Wasserstein distance as a measure. In other terms, we will be focusing on $\mathcal{W}(\tilde{h}_1(t), \tilde{h}_2(t))$ for every $t \geq 0$. The results in the previous sections will provide the necessary background for obtaining the desired bounds. Before displaying these, we state a technical result. To this end, we recall that if ν is a measure on $\mathbb{X}$ and $g : \mathbb{X} \to \mathbb{Y}$ is a measurable function, the pushforward measure $g \# \nu$ on $\mathbb{Y}$ is defined by $(g \# \nu)(A) = \nu(g^{-1}(A))$.

Lemma 10. *Let $\tilde{\mu}$ be a CRM with intensity measure ν and let $f : \mathbb{X} \to \mathbb{R}^+$ be a measurable function. Then the random measure $\tilde{\mu}_f(dy) = f(y)\,\tilde{\mu}(dy)$ is a CRM with Lévy intensity equal to the pushforward measure $\nu_f = p_f \# \nu$ where $p_f(s, y) = (s\, f(y), y)$. Thus for every $A \in \mathcal{X}$,*

$$\int_{\mathbb{R}^+ \times A} s\, \nu_f(ds, dx) = \int_{\mathbb{R}^+ \times A} f(y)\, s\, \nu(ds, dy). \tag{2.9}$$

When $\nu(ds, dy) = \nu(s, y)\, ds\, \alpha(dy)$, with a change of variable Lemma 10 ensures that

$$\nu_f(ds, dy) = \frac{1}{f(s)}\, \nu\Big(\frac{s}{f(y)}, y\Big) ds\, \alpha(dy).$$

Thus, we will use the notation $\nu_f(ds, dy) = \frac{1}{f(s)}\nu(d\frac{s}{f(y)}, dy)$. The relevance of this change of measure result is apparent since the prior specification in (2.8) involves a multiplicative structure with the kernel and the mixing CRM. The following example deals with the gamma case.

Example 2. Consider $\tilde{\mu} \sim \mathrm{Ga}(b, \alpha)$ and a generic kernel k. Then the random measures defined by $\tilde{\mu}_{k(t|\cdot)}(dy) = k(t \,|\, y)\tilde{\mu}(dy)$ are CRMs with Lévy intensity

$$\nu_{k(t|\cdot)}(ds, dy) = \frac{e^{-\frac{sb}{k(t\,|\,y)}}}{s}\, \mathbb{1}_{(0,+\infty)}(s)\, ds\, \alpha(dy).$$

Thus $\tilde{\mu}_{k(t|\cdot)}$ is an *extended gamma* CRM with scale function $\beta(y) = \frac{k(t\,|\,y)}{b}$ and base measure α. Extended gamma CRMs are easily shown to be infinitely active.

Lemma 10 ensures that marginally the hazard process in (2.8) satisfies $\tilde{h}(t) \overset{\mathrm{d}}{=} \tilde{\mu}_{k(t|\cdot)}(\mathbb{X})$, where $\tilde{\mu}_{k(t|\cdot)}$ is a CRM. In order to bound the Wasserstein distance between marginal hazards we may thus apply the results of Theorem 5 with $A = \mathbb{X}$. By (2.9), $\tilde{\mu}_{k(t|\cdot)}$ has finite total mean and it is infinitely active if, respectively,

$$\int_{\mathbb{R}^+ \times \mathbb{X}} k(t \,|\, y)\, s\, \nu(ds, dy) < +\infty, \tag{2.10}$$

$$\int_{[0,\epsilon] \times A} \frac{1}{k(t \,|\, y)} \nu\Big(d\frac{s}{k(t \,|\, y)}, dy\Big) = +\infty, \tag{2.11}$$

for every $\epsilon \geq 0$, $A \in \mathcal{X}$. If ν is infinitely active, (2.11) holds.

Theorem 11. *Let $\tilde{h}_1 = \{\tilde{h}_1(t) \,|\, t \geq 0\}$ and $\tilde{h}_2 = \{\tilde{h}_2(t) \,|\, t \geq 0\}$ be random hazard rates as in (2.8) with associated infinitely active CRMs $\tilde{\mu}_i$, Lévy intensity ν_i, and kernel k_i that satisfy (2.7) and (2.10), for $i = 1, 2$. Then the Wasserstein distance between the marginal hazard rates is finite and for every $t \geq 0$,*

$$g_\ell(t) \leq \mathcal{W}(\tilde{h}_1(t), \tilde{h}_2(t)) \leq g_u(t),$$

where

$$g_\ell(t) = \left| \int_{\mathbb{R}^+\times\mathbb{X}} k_1(t\,|\,y)\, s\, \nu_1(ds,dy) - \int_{\mathbb{R}^+\times\mathbb{X}} k_2(t\,|\,y)\, s\, \nu_2(ds,dy) \right|;$$
$$g_u(t) = \int_0^{+\infty} \left| \int_{(u,+\infty)\times\mathbb{X}} \frac{1}{k_1(t\,|\,y)} \nu_1\Big(d\frac{s}{k_1(t\,|\,y)}, dy\Big) - \frac{1}{k_2(t\,|\,y)} \nu_2\Big(d\frac{s}{k_2(t\,|\,y)}, dy\Big) \right| du.$$

In particular if there exists a dominating measure η such that the Radon–Nikodym derivatives $\nu_i(s,y)$ satisfy, for $i \neq j$ in $\{1,2\}$,

$$\frac{1}{k_i(t\,|\,y)} \nu_i\Big(\frac{s}{k_i(t\,|\,y)}, y\Big) \leq \frac{1}{k_j(t\,|\,y)} \nu_j\Big(\frac{s}{k_j(t\,|\,y)}, y\Big)$$

for all $(s,y) \in \mathbb{R}^+ \times \mathbb{X}$, then

$$\mathcal{W}(\tilde{h}_1(t), \tilde{h}_2(t)) = \left| \int_{\mathbb{R}^+\times\mathbb{X}} k_1(t\,|\,y)\, s\, \nu_1(ds,dy) - \int_{\mathbb{R}^+\times\mathbb{X}} k_2(t\,|\,y)\, s\, \nu_2(ds,dy) \right|.$$

2.4.2 Bounds for survival functions

The bounds we derived for the hazard rates translate into bounds for the corresponding survival functions and these are of great interest since one typically targets estimation of functionals of the survival function. We proceed by first deriving bounds for the corresponding cumulative hazards processes $\tilde{H} = \{\tilde{H}(t)\,|\,t \geq 0\}$, defined by

$$\tilde{H}(t) = \int_0^t \tilde{h}(u)\, du = \int_{\mathbb{X}} K(t\,|\,y)\, \tilde{\mu}(dy), \tag{2.12}$$

where $K(t\,|\,y) = \int_0^t k(u\,|\,y)\, du$ is the *cumulative kernel*. Thus, the cumulative hazards can be treated as a kernel mixture as well, and an analogue of Theorem 11 is available.

Theorem 12. *Let $\tilde{H}_1 = \{\tilde{H}_1(t)\,|\,t \geq 0\}$ and $\tilde{H}_2 = \{\tilde{H}_2(t)\,|\,t \geq 0\}$ be two random cumulative hazards as in (2.12) with associated infinitely active CRMs $\tilde{\mu}_i$, with Lévy intensity ν_i, and kernel k_i that satisfy (2.7), for $i = 1,2$. If the cumulative kernels $K_i(t\,|\,y) = \int_0^t k_i(u\,|\,y)du$ satisfy (2.10), the Wasserstein distance between the marginal cumulative hazards is finite and for every $t \geq 0$,*

$$g_\ell(t) \leq \mathcal{W}(\tilde{H}_1(t), \tilde{H}_2(t)) \leq g_u(t),$$

where

$$g_\ell(t) = \left| \int_{\mathbb{R}^+\times\mathbb{X}} K_1(t\,|\,y)\, s\, \nu_1(ds,dy) - K_2(t\,|\,y)\, s\, \nu_2(ds,dy) \right|;$$
$$g_u(t) = \int_0^{+\infty} \left| \int_{(u,+\infty)\times\mathbb{X}} \frac{1}{K_1(t\,|\,y)} \nu_1\Big(d\frac{s}{K_1(t\,|\,y)}, dy\Big) - \frac{1}{K_2(t\,|\,y)} \nu_2\Big(d\frac{s}{K_2(t\,|\,y)}, dy\Big) \right| du.$$

In particular if there exists a dominating measure η such that the Radon–Nikodym derivatives $\nu_i(s,y)$ satisfy, for $i \neq j$ in $\{1,2\}$,

$$\frac{1}{K_i(t\,|\,y)}\nu_h\Big(\frac{s}{K_i(t\,|\,y)},y\Big) \leq \frac{1}{K_j(t\,|\,y)}\nu_j\Big(\frac{s}{K_j(t\,|\,y)},y\Big)$$

for all $s,y \in \mathbb{R}^+ \times \mathcal{X}$, then

$$\mathcal{W}(\tilde{H}_1(t),\tilde{H}_2(t)) = \Big|\int_{\mathbb{R}^+\times\mathbb{X}} K_1(t\,|\,y)\,s\,\nu_1(ds,dy) - \int_{\mathbb{R}^+\times\mathbb{X}} K_2(t\,|\,y)\,s\,\nu_2(ds,dy)\Big|.$$

The bounds for the distance between cumulative hazards in Theorem 12 are also useful to identify a similar result for the survival function process $\tilde{S} = \{\tilde{S}(t)\,|\,t \geq 0\}$ defined by

$$t \mapsto \tilde{S}(t) = e^{-\tilde{H}(t)} = \exp\Big\{-\int_{\mathbb{X}} K(t\,|\,y)\,\tilde{\mu}(dy)\Big\}. \tag{2.13}$$

Theorem 13. *Let $\tilde{H}_1$ and $\tilde{H}_2$ be as in* Theorem 12 *with survival process $\tilde{S}_i$ as in* (2.13)*, for $i = 1,2$. Then for every $t \geq 0$,*

$$g_\ell(t) \leq \mathcal{W}(\tilde{S}_1(t),\tilde{S}_2(t)) \leq \min\{g_{u,1}(t), g_{u,2}(t)\},$$

where

$$g_\ell(t) = \big|\mathbb{E}\big(e^{-\tilde{H}_1(t)}\big) - \mathbb{E}\big(e^{-\tilde{H}_2(t)}\big)\big|, \qquad g_{u,1}(t) = 1 - e^{-\mathcal{W}(\tilde{H}_1(t),\tilde{H}_2(t))},$$

$$g_{u,2}(t) = \mathbb{E}\big(e^{-\tilde{H}_1(t)}\big) + \mathbb{E}\big(e^{-\tilde{H}_2(t)}\big) - \big(e^{-\mathbb{E}(\tilde{H}_1(t))} + e^{-\mathbb{E}(\tilde{H}_2(t))}\big)e^{-\mathcal{W}(\tilde{H}_1(t),\tilde{H}_2(t))}.$$

2.4.3 Examples

We now apply these results on kernels of the type of Dykstra & Laud (1981), $k(t|y) = \beta(y)\mathbb{1}_{[0,t]}(y)$, which is a popular choice when one wants to model increasing hazards. In this setting $\mathbb{X} = [0,+\infty)$. For simplicity we will restrict our attention to constant functions $\beta(s) = \beta$, which is a common choice in applications (Dykstra & Laud, 1981; Laud et al., 1996), and gamma CRMs with the same base measure α. In this scenario α may also be an infinite measure, though it must be boundedly finite. We will consider the Lebesgue measure on the positive real axis, $\mathscr{L}^+(ds) = \mathbb{1}_{[0,+\infty)}(s)\,ds$, which is the base measure proposed in the original paper of Dykstra & Laud (1981) and meets the conditions of Theorem 11.

Example 3. Let $\tilde{\mu}_i \sim \mathrm{Ga}(b_i, \mathscr{L}^+)$ and let $k_i(t\,|\,y) = \beta_i\mathbb{1}_{[0,t]}(y)$, with $b_i, \beta_i > 0$, for $i = 1,2$. If $\tilde{h}_1$ and $\tilde{h}_2$ are the corresponding hazard rate mixtures, then

$$\mathcal{W}(\tilde{h}_1(t),\tilde{h}_2(t)) = t\left|\frac{\beta_1}{b_1} - \frac{\beta_2}{b_2}\right|.$$

Proof. The general expression for $\nu_{k(t|\cdot)}$ was derived in Example 2. With our choices,

$$\frac{1}{k_i(t\,|\,y)}\nu_i\Big(d\frac{s}{k_i(t\,|\,y)}, dy\Big) = \frac{e^{-\frac{s\,b_i}{\beta_i}}}{s}\,\mathbb{1}_{(0,+\infty)}(s)\,\mathbb{1}_{[0,t]}(y)\,ds\,dy, \tag{2.14}$$

which corresponds to the Lévy intensity of a gamma CRM of parameter b_i/β_i and the restriction of $\mathscr{L}^+$ to $[0,t]$ as base measure. Since the Lebesgue measure on a bounded set is finite, as observed in Proposition 8, (2.14) is infinitely active and has finite mean. Thus condition (2.10) holds. In order to check condition (2.7) on the expected cumulative hazards we first observe that for every $t > 0$,

$$\mathbb{E}(\tilde{h}_i(t)) = \int_{\mathbb{R}^+\times\mathbb{R}^+} \beta_i\, e^{-s\,b_i}\,\mathbb{1}_{[0,t]}(y)\,ds\,dy = t\,\frac{\beta_i}{b_i}.$$

Thus $\int_0^t \mathbb{E}(\tilde{h}_i(s))\,ds = t^2\,\frac{\beta_i}{2\,b_i}$ diverges as $t \to +\infty$, and condition (2.7) holds. The results in Theorem 11 apply and since the densities of (2.14) are ordered, we easily derive the expression for $\mathcal{W}(\tilde{h}_1(t), \tilde{h}_2(t))$ from the expression of $\mathbb{E}(\tilde{h}_i(t))$, in accordance with the results in Proposition 8 for gamma CRMs.

□

The choice of the kernel allows for great flexibility and usually depends on the type of experiment one is considering. For example, if we are dealing with the failure of objects whose material wears out in time, the assumption of increasing hazard rates appears to be the most plausible. Besides the kernel of Dykstra & Laud (1981), which leads to almost surely increasing hazard rates, one can resort to other options such as:

(1) Rectangular kernel with threshold τ: $k(t\,|\,x) = \mathbb{1}_{[t-\tau,t+\tau]}(x)$;

(2) Bathtub or U-shaped kernel with minimum $\eta > 0$: $k(t\,|\,x) = \mathbb{1}_{[0,|t-\eta|]}(x)$;

(3) Ornstein-Uhlenbeck kernel with $g > 0$: $k(t\,|\,x) = \sqrt{2g}\,e^{-g(t-x)}\mathbb{1}_{[0,t]}(x)$;

(4) Exponential kernel: $k(t\,|\,x) = e^{-tx}$.

More details can be found in Lo & Weng (1989); James (2003); De Blasi et al. (2009). The choice of the kernel is typically dictated by the type of data one is examining. As for the choice of random measure, this may be dictated by specific inferential properties but it is usually motivated by analytical tractability and prior flexibility. In this regards, the gamma CRM is a popular alternative. We thus pick one of the kernels above and focus on gamma kernel mixtures. One is, then, left with the choice of the parameter b, which heuristically quantifies the prior belief on the steepness of the hazard. Given these specifications, one may be interested in quantifying the discrepancy induced by it on the corresponding hazards. Before proceeding, we underline how the same reasoning could be applied to the base measure, but for simplicity we consider gamma CRMs with a given shared base measure. We point out that in all cases we achieve the exact expression for the Wasserstein distance between the hazard rates.

Example 4. Let $\tilde{\mu}_i \sim \mathrm{Ga}(b_i, \mathscr{L}^+)$, with $b_i > 0$, for $i = 1, 2$. Let $k_1 = k_2 = k$ be one of the kernels (1)–(4) above. Then the Wasserstein distances between the corresponding hazard rates mixtures equal

(a) $\mathcal{W}(\tilde{h}_1(t), \tilde{h}_2(t)) = (2\tau - (\tau - t)^+)\, |b_1^{-1} - b_2^{-1}|$;

(b) $\mathcal{W}(\tilde{h}_1(t), \tilde{h}_2(t)) = |t - \eta|\, |b_1^{-1} - b_2^{-1}|$;

(c) $\mathcal{W}(\tilde{h}_1(t), \tilde{h}_2(t)) = g\sqrt{2g}(1 - e^{-gt})\, |b_1^{-1} - b_2^{-1}|$;

(d) $\mathcal{W}(\tilde{h}_1(t), \tilde{h}_2(t)) = t^{-1}(1 - e^{-t^2})\, |b_1^{-1} - b_2^{-1}|$;

where $f^+ = \max(f, 0)$ for any measurable function f with values in $\mathbb{R}$.

Proof. Kernel (a) is very similar to the one in Example 3. The Lévy intensity

$$\frac{1}{k_i(t \,|\, y)} \nu_i\Big(d\frac{s}{k_i(t \,|\, y)}, dy\Big) = \frac{e^{-s\, b_i}}{s}\, \mathbb{1}_{(0,+\infty)}(s)\, \mathbb{1}_{[0\wedge(t-\tau), t+\tau]}(y)\, ds\, dy$$

is the one of a gamma CRM with parameter b and Lebesgue base measure on $[0 \wedge (t - \tau), t + \tau]$. Since the corresponding densities are ordered, the exact Wasserstein distance is available and coincides with (a′). The same is true for kernel (b). With kernel (c) one has

$$\frac{1}{k_i(t \,|\, y)} \nu_i\Big(d\frac{s}{k_i(t \,|\, y)}, dy\Big) = \frac{1}{s} \exp\Big\{ - \frac{s b_i}{\sqrt{2g}} e^{g(t-y)} \Big\} \mathbb{1}_{(0,+\infty)}(s)\, \mathbb{1}_{[0,t]}(y)\, dy\, ds.$$

The corresponding densities are ordered, thus if the conditions of Theorem 11 hold we only need to evaluate the expected value of the hazards to derive the exact Wasserstein distance:

$$\begin{aligned} \mathbb{E}(\tilde{h}_i(t)) &= \int_{\mathbb{R}^+ \times \mathbb{R}} \sqrt{2g}\, e^{-g(t-y)} e^{-b_i s}\, \mathbb{1}_{[0,t]}(y)\, dy\, ds \\ &= \int_0^t -\frac{\sqrt{2g}}{b_i} e^{-g(t-y)} dy = \sqrt{\frac{2}{g}}(1 - e^{gt}) \frac{1}{b_i}. \end{aligned}$$

This also proves the finite mean condition (2.10). Since $\int_0^t (1 - e^{gs}) ds = \frac{1 - e^{gt}}{g} + t$ diverges as $t \to +\infty$, also condition (2.7) holds.
Finally for kernel (d),

$$\frac{1}{k_i(t \,|\, y)} \nu_i\Big(d\frac{s}{k_i(t \,|\, y)}, dy\Big) = \frac{1}{s} \exp\big\{ - s b e^{ty} \big\} \mathbb{1}_{(0,+\infty)}(s)\, \mathbb{1}_{[0,t]}(y)\, dy\, ds.$$

The mean hazard rates are

$$\mathbb{E}(\tilde{h}_i(t)) = \int_0^t \frac{e^{-ty}}{b_i} dy = \frac{1 - e^{-t^2}}{b_i t},$$

and thus condition (2.10) holds. Moreover,

$$\int_0^t \frac{1-e^{-s^2}}{s} ds = \frac{\gamma}{2} + \frac{E_1(t^2)}{2} + \log(t),$$

where γ is the Euler gamma constant. This quantity diverges as $\log(t)$ for $t \to +\infty$ and thus condition (2.7) holds. We conclude as in the previous cases.

□

Next, we apply the bounds on cumulative hazards and survival functions of Theorem 12 and Theorem 13 to the case where mixtures of gamma CRMs are used, as in Example 3.

Example 5. Consider the prior specification of Example 3. Denote by $\tilde{H}_i$ the corresponding cumulative process (2.12) and by $\tilde{S}_i$ the corresponding survival process (2.13), for $i = 1, 2$. Then for every $t \geq 0$,

$$\mathcal{W}(\tilde{H}_1(t), \tilde{H}_2(t)) = \frac{t^2}{2}\left|\frac{\beta_1}{b_1} - \frac{\beta_2}{b_2}\right|, \tag{2.15}$$

$$g_\ell(\boldsymbol{b}, t) \leq \mathcal{W}(\tilde{S}_1(t), \tilde{S}_2(t)) \leq g_{u,1}(\boldsymbol{b}, t) \wedge g_{u,2}(\boldsymbol{b}, t), \tag{2.16}$$

where

$$g_\ell(\boldsymbol{b}, t) = e^t \left| \left(\frac{b_1}{b_1+\beta_1 t}\right)^{\frac{b_1+\beta_1 t}{\beta_1}} - \left(\frac{b_2}{b_2+\beta_2 t}\right)^{\frac{b_2+\beta_2 t}{\beta_2}} \right|,$$

$$g_{u,1}(\boldsymbol{b}, t) = 1 - e^{-\frac{t^2}{2}\left|\frac{\beta_1}{b_1} - \frac{\beta_2}{b_2}\right|},$$

$$g_{u,2}(\boldsymbol{b}, t) = e^t \left(\left(\frac{b_1}{b_1+\beta_1 t}\right)^{\frac{b_1+\beta_1 t}{\beta_1}} + \left(\frac{b_2}{b_2+\beta_2 t}\right)^{\frac{b_2+\beta_2 t}{\beta_2}} \right) - \left(e^{-\frac{t^2\beta_1}{2b_1}} + e^{-\frac{t^2\beta_2}{2b_2}} \right) e^{-\frac{t^2}{2}\left|\frac{\beta_1}{b_1} - \frac{\beta_2}{b_2}\right|}.$$

Proof. Since the Lévy densities of

$$\frac{1}{K_i(t \,|\, y)} \nu_i\left(d\frac{s}{K_i(t \,|\, y)}, dy\right) = \frac{1}{s} \exp\left\{ -\frac{s b_i}{t-y} \right\} \mathbb{1}_{(0,+\infty)}(s)\, \mathbb{1}_{[0,t]}(y)\, dy\, ds$$

are ordered, if the conditions of Theorem 12 hold the expression for the Wasserstein distance between the cumulative hazards easily derives from

$$\mathbb{E}(\tilde{H}_i(t)) = \int_{\mathbb{R}^+ \times [0,+\infty)} K_i(t \,|\, y)\, s\, \nu_i(ds, dy) = \int_{\mathbb{R}^+} \int_0^t \beta_i (t-y)\, e^{-s b_i} ds\, dy = \frac{t^2 \beta_i}{2 b_i}. \tag{2.17}$$

Now, condition (2.7) on the kernels has already been checked in Example 3. Moreover, (2.17) proves condition (2.10) on the finite mean.
As for the Wasserstein distance between the survival functions, in order to apply Theorem 13 it suffices to evaluate the mean of the survival functions. This is easily done thanks to the properties of the Laplace functional of a CRM. Specifically, $\mathbb{E}\left(e^{-\int_{\mathbb{R}} K(t \,|\, y)\, \tilde{\mu}_i(dy)}\right)$ is equal to

$$\exp\left\{ -\int_{\mathbb{R}^+ \times [0,+\infty)} \left(1 - e^{s\, K_i(t \,|\, y)}\right) \nu_i(ds, dy) \right\} = \left(\frac{b_i}{b_i + \beta_i t}\right)^{\frac{b_i+\beta_i t}{\beta_i}}.$$

□

In Figure 2.2 a graphical representation of the upper and lower bounds for the Wasserstein distance between the corresponding survival functions is given. In particular, the distance between the survival functions lies in the gray area in the figure. The first upper bound $g_{u,1}$ appears to be tighter for small times, while the second $g_{u,2}$ is more informative as time increases. This depends on the fact that in the first case we are using the bound $e^{-\tilde{H}_1(t)\wedge\tilde{H}_2(t)} \leq 1$, which is effective for small values of the cumulative hazard function, i.e. for small times, while in the second one we are using $e^{-\tilde{H}_1(t)\wedge\tilde{H}_1(t)} \leq e^{-\tilde{H}_1(t)} + e^{-\tilde{H}_2(t)}$, which is effective for large values of the cumulative hazard function, i.e. for large t. Moreover, we point out that the Wasserstein distance between survival functions is considerably smaller than the one between the hazard rates, which is what we expect from a modeling perspective.

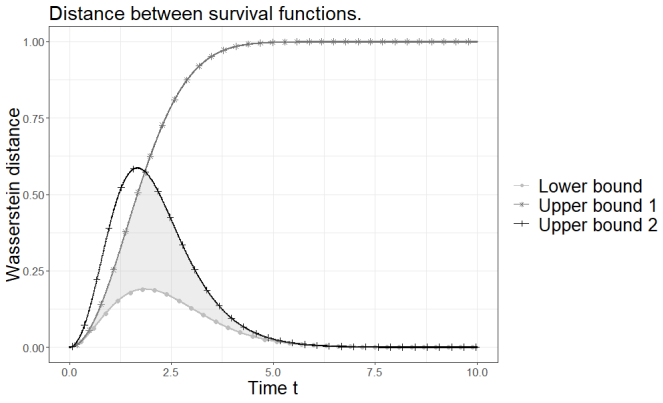

Figure 2.2: Theoretical upper and lower bounds for the Wasserstein distance between marginals of the random survival functions in Example 5 with $b_1 = 1$, $\beta_1 = 1$, $b_2 = 2$ and $\beta_2 = 1$.

2.5 Posterior sampling scheme

The techniques we have developed in the previous sections may be fruitfully applied to evaluate approximation errors in posterior sampling schemes. In this section we will focus on the gamma kernel mixture by Dykstra & Laud (1981) and rely on the posterior analysis by James (2005). Even when the prior hazards are modeled as a gamma process (i.e. constant β), conditionally on the data and a set of latent variables, the non–atomic part of the posterior hazards is an extended gamma process. There are many available methods in the literature to sample from an extended gamma process, as the finite dimensional approximation by Ishwaran & James (2004), the inverse Lévy methods of Ferguson & Klass (1972) and Wolpert & Ickstadt (1998), and the series representation of Bondesson (1982), which serves as a basis for the algorithm in Laud et al. (1996). Other available series representations can be found in Rosiński (2001). Recently, Al Masry et al. (2017) proposed a new algorithm based on a discretization of the scale function: in such case the extended gamma process can be approximated by a sum of gamma

random increments. The construction of the discretization is not always simple but, when possible, it allows for a precise quantification of the approximation error. In Al Masry et al. (2017) the error is quantified through a bound on the L^2 distance. Here, we build the discrete approximation of the scale function of the posterior hazards corresponding to a gamma process prior and use the Wasserstein distance to quantify the approximation error between the induced hazard rates. Moreover, since one is usually interested in the cumulative hazards or, more often, in the survival function, we provide an estimate for their approximations as well, which provides a novel and meaningful guide for fixing the approximation error in the algorithm.

We first recall the posterior characterization of mixture hazard rates models, with censored data, as achieved by James (2005). This result is suited to our case, since it concerns CRM–driven mixtures under a multiplicative intensity model. In order to provide a summary description of the posterior distribution, henceforth $T_1, \dots, T_n$ are random elements from an exchangeable sequence as in (2.6) with Π being the law of a random probability measure with hazard rate $\tilde{h}$ as in (2.8). Furthermore, if $n_e = \sum_{i=1}^{n} \Delta_i$ is the number of exact observations in the sample, we may assume without loss of generality that $\Delta_1 = \cdots = \Delta_{n_e} = 1$ and, hence, the last $n - n_e$ observations are censored. The data are, then, given by $\{(x_j, \Delta_j)\}_{j=1}^{n}$. A representation of the likelihood function that is convenient for Bayesian computations is obtained by relying on a suitable augmentation that involves a collection of latent variables $Y_1, \dots, Y_{n_e}$ corresponding to the exact observations. Hence, the augmented likelihood is given by

$$\begin{aligned}\mathcal{L}(\tilde{\mu}; \boldsymbol{x}, \boldsymbol{y}) &= e^{-\int_{\mathbb{X}} K_n(y)\tilde{\mu}(dy)} \prod_{j=1}^{n_e} \tilde{\mu}(dy_j) k(x_j \,|\, y_j) \\ &= e^{-\int_{\mathbb{X}} K_n(y)\tilde{\mu}(dy)} \prod_{h=1}^{k} \tilde{\mu}(dy_h^*)^{n_h} \prod_{i \in C_h} k(x_i \,|\, y_h^*),\end{aligned}$$

where $\boldsymbol{x} = (x_1, \dots, x_n)$, $y_1^*, \cdots, y_k^*$ are the $k \leq n_e$ distinct values in $\boldsymbol{y} = (y_1, \dots, y_{n_e})$, $C_j = \{r : y_r = y_j^*\}$ and $n_j = \text{card}(C_j)$. The function K_n is interpretable as a kernel for the cumulative hazards and, in general, accounts for different forms of censoring. For simplicity we henceforth focus on the case of right-censored observations and this yields

$$K_n(y) = \sum_{j=1}^{n} \int_0^{x_j} k(u \,|\, y)\, du. \tag{2.18}$$

The posterior characterization we rely on is as follows.

Theorem 14 (James (2005))**.** *Let $T_1, \dots, T_n$ be random elements from an exchangeable sequence as in (2.6), with Π being the law of a random probability measure with hazard rate $\tilde{h}$ as in (2.8). Conditional on the observed data $\boldsymbol{x} = (x_1, \cdots, x_n)$ and latent variables $\boldsymbol{y} = (y_1, \cdots, y_{n_e})$, the posterior distribution of $\tilde{\mu}$ equals in distribution*

$$\tilde{\mu}^* \stackrel{d}{=} \tilde{\mu}_c^* + \sum_{h=1}^{k} J_h \delta_{y_h^*}, \tag{2.19}$$

where $\tilde{\mu}_c^*$ *is a CRM without fixed jump points and with intensity*

$$\nu^*(ds, dy) = e^{-sK_n(y)}\nu(ds, dy) = e^{-sK_n(y)}\rho_y(ds)\,\eta(dy),$$

while $J_1, \cdots, J_k$ *are mutually independent and independent from* $\tilde{\mu}_c^*$*. For* $h = 1, \ldots, k$*, the generic* h*–th jump* J_h *has distribution*

$$G_h(ds) \propto s^{n_h} e^{-sK_n(y_h^*)}\,\rho_{y_h^*}(ds). \tag{2.20}$$

In the rest of the section we focus on the case $\mathbb{X} = \mathbb{R}$, $k(t|y) = \beta \mathbb{1}_{[0,t]}(y)$ and μ gamma CRM with rate parameter b and base measure α, which is a typical choice in applications. Thus the non–atomic posterior CRM $\tilde{\mu}_c^*$ has Lévy measure

$$\nu^*(ds, dy) = \frac{e^{-s(b+\beta\sum_{i=1}^n (y-x_i)^+)}}{s}\,\mathbb{1}_{(0,+\infty)}(s)\,ds\,\alpha(dy).$$

It follows that $\tilde{\mu}_c^*$ is an extended gamma CRM with base measure α and scale function $1/(b + \beta\sum_{i=1}^n (y - x_i)^+)$. The non–atomic posterior hazards are an extended gamma process and can thus be written as

$$\tilde{h}^*(t) \overset{\mathrm{d}}{=} \int_0^t \beta^*(s)\,\tilde{\mu}(ds), \tag{2.21}$$

where $\beta^*(y) = \beta/(b + \beta\sum_{i=1}^n (y - x_i)^+)$ and $\tilde{\mu}$ is a gamma CRM with parameter 1 and base measure α.
Consider an interval of interest $[0, T]$, which can be thought as the initial and final time of the study, so that $0 < x_1 \leq \cdots \leq x_n < T$. The algorithm proposed by Al Masry et al. (2017) to sample from $\{\tilde{h}^*(t)\,|\,t \in [0, T]\}$ is based on a piecewise constant approximation of β^* on the interval $[0, T]$. If $\beta^\epsilon(y) = \sum_{h=0}^{n(\epsilon)} \beta_h \mathbb{1}_{(t_h, t_{h+1}]}(y)$, then for every $t \geq 0$,

$$\tilde{h}^\epsilon(t) = \int_0^t \beta^\epsilon(s)\,\tilde{\mu}(ds) = \sum_{h=1}^{n_t} \beta_h\,\tilde{\mu}(t_h, t_{h+1}] + \beta_{n_t+1}\,\tilde{\mu}(t_{n_t}, t], \tag{2.22}$$

where n_t is such that $t_{n_t} \leq t \leq t_{n_t+1}$. We observe that the increments $\delta_h = \beta_h\,\tilde{\mu}(t_h, t_{h+1}]$ have a gamma distribution with scale β_h and shape $\alpha(t_h, t_{h+1})$. If the points$\{t_h\,|\,h = 1, \ldots, n(\epsilon)\}$ are dense in the interval $[0, T]$ as $n(\epsilon) \to +\infty$, one samples directly from a sum of gamma random variables $\sum_{t_h \leq t} \delta_h$.
In order to apply this algorithm we need to build an approximating strictly positive piecewise constant function $\beta^\epsilon : [0, T] \to (0, +\infty)$ and find a reasonable criterion to fix the approximation error. We will build β^ϵ by discretizing the reciprocal of β^*, namely $\gamma(y) = b\,\beta^{-1} + \sum_{i=1}^n (y - x_i)^+$. Consider the points $t_0 \leq t_1 \leq \cdots \leq t_{n(\epsilon)-1} = x_n \leq t_{n(\epsilon)} = T$ defined by

$$t_{j+\sum_{i=0}^{j-1}[\epsilon^{-1}(n-i)(x_{i+1}-x_i)]+k} = x_j + \frac{k\,\epsilon}{n-j},$$

for every $j = 0, \dots, n-1$ and $k = 0, \dots, \big[(n-j)(x_{j+1} - x_j)\epsilon^{-1}\big]$, where $x_0 = 0$ and $[\cdot]$ denotes the integer part, so that $n(\epsilon) = n + 1 + \sum_{i=0}^{n-1} \big[(n-i)(x_{i+1} - x_i)\epsilon^{-1}\big]$. We observe that

$$\gamma\Big(t_{j+\sum_{i=0}^{j-1}[\epsilon^{-1}(n-i)(x_{i+1}-x_i)]+k}\Big) = \frac{b}{\beta} + \sum_{i=j+1}^{n} x_i - (n-j)x_j - k\,\epsilon.$$

Next we set $\gamma_h = \gamma(t_{h+1})$ and define

$$\gamma^\epsilon(y) = \sum_{h=0}^{n(\epsilon)} \gamma_i \mathbb{1}_{(t_h, t_{h+1}]}(y). \tag{2.23}$$

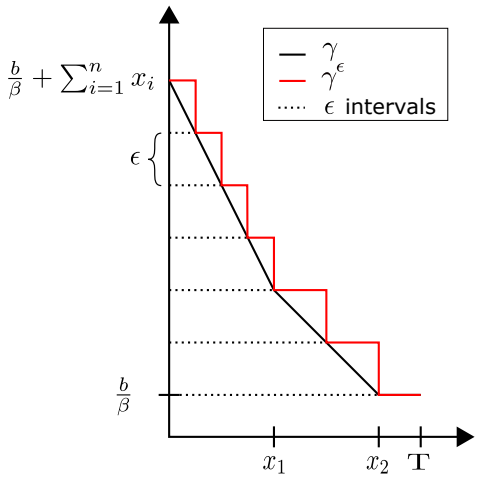

Figure 2.3: Piecewise constant approximation γ^ϵ of the function γ on the interval $[0, T]$.

Theorem 15. *The function defined in (2.23) is piecewise constant and satisfies*

$$\sup_{y \in ([0,T]} |\gamma(y) - \gamma^\epsilon(y)| \le \epsilon.$$

Moreover, $\gamma^\epsilon(y) \ge \gamma(y)$ for every $y \in [0, T]$.

From this theorem one easily deduces a very simple uniform bound for the discrepancy between β^ϵ and β^*.

Corollary 16. *If $\beta^\epsilon(y) = \frac{1}{\gamma^\epsilon(y)} = \sum_{h=0}^{n(\epsilon)} \frac{1}{\gamma_h} \mathbb{1}_{(t_h, t_{h+1}]}(y)$, then one has*

$$\sup_{y \in (0,T]} |\beta^*(y) - \beta^\epsilon(y)| \le \frac{\beta^2}{b^2}\epsilon. \tag{2.24}$$

For a given ϵ, these results provide a constructive rule for determining an approximation of β^*, and hence an approximating hazard $\tilde{h}^\epsilon$. One may then wonder which value of ϵ should be specified if we wish to achieve a prescribed error of approximation for the

posterior hazards and survivals. This is achieved in the next result, where we propose three different rules based on the Wasserstein distance between the hazards, cumulative hazards and the survival functions respectively. The result is stated for the hypotheses of Example 3, $\alpha = \mathscr{L}^+$, but can be easily adapted to any base measure.

Theorem 17. *Consider the hypotheses of Theorem 14 with $\tilde{\mu} \sim Ga(b, \mathscr{L}^+)$ and $k(y|t) = \beta \mathbb{1}_{(0,t]}(y)$. Let $\tilde{h}^* = \{\tilde{h}^*(t) \,|\, t \in [0,T]\}$ be the non–atomic posterior hazard rates process (2.21), and let $\tilde{h}^\epsilon = \{\tilde{h}^\epsilon(t) \,|\, t \in [0,T]\}$ be its approximation (2.22). If $\tilde{H}^*$, $\tilde{H}^\epsilon$, $\tilde{S}^*$, $\tilde{S}^\epsilon$ denote their respective cumulative hazards and survival functions processes, then*

$$\sup_{t\in(0,T]} \mathcal{W}(\tilde{h}^*(t), \tilde{h}^\epsilon(t)) \leq \epsilon \frac{\beta^2}{b^2} T;$$

$$\sup_{t\in(0,T]} \mathcal{W}(\tilde{H}^*(t), \tilde{H}^\epsilon(t)) \leq \epsilon \frac{\beta^2}{2b^2} T^2;$$

$$\sup_{t\in(0,T]} \mathcal{W}(\tilde{S}^*(t), \tilde{S}^\epsilon(t)) \leq 1 - \exp\left\{ - \epsilon \frac{\beta^2}{2b^2} T^2 \right\}.$$

Theorem 17 determines three different cut–off rules for approximating the posterior estimate coming from the model of Dykstra & Laud (1981). This is crucial for practitioners that are interested in implementing a nonparametric model for nondecreasing hazards and plan on leveraging prior information on the shape of the hazards. Previous studies or expert opinions may be included in the choice of the hyperparameters β and b. To this end, it is useful to observe that $\mathbb{E}(\tilde{h}(t)) = \beta/bt$ and $\mathrm{Var}(\tilde{h}(t) = \beta^2/bt$. On the other hand, the time-interval of interest $[0,T]$ is typically dictated by the length of the experiment. Depending on the study, the main quantity of interest may be the hazard function, the cumulative hazards or, more often, the survival function. One then picks a level of tolerance δ for the quantity of interest and equates it to the corresponding upper bound in Theorem 17, so to find the desired cut–off value ϵ for the algorithm.

2.6 Proofs

2.6.1 Proof of Theorem 5

First of all we state a technical lemma.

Lemma 18. *Let $\tilde{\mu}$ be a CRM with Lévy intensity ν and finite total mean (2.4). Then for every $A \in \mathcal{X}$,*

$$\lim_{\epsilon\to 0^+} \epsilon\, \nu([\epsilon, +\infty) \times A) = 0.$$

Proof. For every $\delta > 0$ consider $\epsilon > 0$ such that $\epsilon < \delta$. Then

$$\epsilon\, \nu([\epsilon, +\infty) \times A) = \epsilon \int_\epsilon^\delta \int_A \nu(ds, dy) + \epsilon \int_\delta^{+\infty} \int_A \nu(ds, dy).$$

The second integral is finite by (2.1), thus $\epsilon \int_\delta^{+\infty} \int_A \nu(ds, dy) \to 0$ as $\epsilon \to 0$. As for the first one, this can be bounded by

$$\epsilon \int_\epsilon^\delta \int_A \nu(ds, dy) \leq \int_\epsilon^\delta \int_A s\, \nu(ds, dy).$$

Since the integrand is integrable in $[0, \delta]$ thanks to the finite total mean condition (2.4), by the dominated convergence theorem,

$$\limsup_{\epsilon \to 0} \epsilon \int_\epsilon^\delta \int_A \nu(ds, dy) \leq \int_0^\delta \int_A s\, \nu(ds, dy).$$

Since this is true for every $\delta > 0$, $\epsilon \int_\epsilon^\delta \int_A \nu(ds, dy) \to 0$ as $\epsilon \to 0$ by the absolute continuity of the integral. □

We now prove the results in Theorem 5. The lower bound $g_\ell(A)$ is easily achieved by (2.3) and by Campbell's theorem applied to the underlying Poisson random measures with respect to the measurable function $f(s, x) = s\, \mathbb{1}_A(x)$, similarly to (2.4). We thus concentrate on the upper bound.
Since the Lévy intensities are diffuse and infinitely active, for every $A \in \mathcal{X}$ and $r > 0$ there exists $\epsilon_{i,r,A} > 0$ such that

$$\nu_i([\epsilon_{i,r,A}, +\infty) \times A) = r, \tag{2.25}$$

for $i = 1, 2$. By denoting with $\mathcal{N}_i$ the Poisson random measure underlying $\tilde{\mu}_i$ as in (1.3),

$$\tilde{\mu}_i(A) \overset{\mathrm{d}}{=} \int_0^{\epsilon_{i,r,A}} \int_A s\, \mathcal{N}_i(ds, dy) + \int_{\epsilon_{i,r,A}}^{+\infty} \int_A s\, \mathcal{N}_i(ds, dy).$$

We use the notation $J^S_{i,r,A} = \int_0^{\epsilon_{i,r,A}} \int_A s\, \mathcal{N}_i(ds, dy)$ for the small jumps and $J^B_{i,r,A} = \int_{\epsilon_{i,r}}^{\infty} \int_A s\, \mathcal{N}_i(ds, dy)$ for the big jumps. The independence of the increments of a Poisson random measure ensures that $J^S_{i,r,A}$ and $J^B_{i,r,A}$ are independent, thus by (A.3)

$$\mathcal{W}(\tilde{\mu}_1(A), \tilde{\mu}_2(A)) \leq \mathcal{W}(J^S_{1,r,A}, J^S_{2,r,A}) + \mathcal{W}(J^B_{1,r,A}, J^B_{2,r,A}).$$

We first show that the small jumps do not play any role in the final bound. By (2.3),

$$\mathcal{W}(J^S_{1,r,A}, J^S_{2,r,A}) \leq \mathbb{E}(J^S_{1,r,A}) + \mathbb{E}(J^S_{2,r,A}).$$

The means $\mathbb{E}(J^S_{i,r,A}) = \int_0^{\epsilon_{i,r,A}} s\, \nu_i(ds \times A)$ are finite by (2.1) and thus go to zero as $r \to +\infty$ by the absolute continuity of the integral. We now focus on the big jumps. By (2.1), these are integrals of Poisson random measures with finite mean measure, $\nu_i(ds, dy)\, \mathbb{1}_{[\epsilon_{i,r,A}, +\infty)}(s)$. Proposition 19.5 in Sato (1999) then ensures that $J^B_{i,r,A}$ has a compound Poisson distribution, so that

$$J^B_{i,r,A} \overset{\mathrm{d}}{=} \sum_{h=1}^{N_{i,r,A}} \xi^h,$$

where $N_{i,r,A}$ is a Poisson random variable with intensity $r = \nu_i([\epsilon_{i,r,A}, +\infty) \times A)$ and $(\xi^h)_h$ are iid random variables, independent from $N_{i,r,A}$, with distribution

$$\rho_{i,r,A}(ds) = \frac{1}{r} \int_A \nu_i(ds, dy)\, \mathbb{1}_{[\epsilon_{i,r,A}, +\infty)}(s). \tag{2.26}$$

Theorem 10 in Mariucci & Reiß (2018) deals with the Wasserstein distance between compound Poisson distributions. Since $J^B_{1,r,A}$ and $J^B_{2,r,A}$ have the same total intensity measure r but different jump distribution $\rho_{i,r,A}$, an immediate adaptation of their result yields

$$\mathcal{W}(J^B_{1,r,A}, J^B_{2,r,A}) \leq r\, \mathcal{W}(\rho_{1,r,A}, \rho_{2,r,A}).$$

By (A.2), $\mathcal{W}(\rho_{1,r,A}, \rho_{2,r,A}) = \int_{-\infty}^{+\infty} |F_{1,r,A}(u) - F_{2,r,A}(u)|\, du$, where $F_{i,r,A}(u)$ is equal to

$$\frac{1}{r}\nu_i([\epsilon_{i,r,A}, u] \times A)\, \mathbb{1}_{[\epsilon_{i,r,A}, +\infty)}(u) = \left(1 - \frac{1}{r}\nu_i((u, +\infty) \times A)\right) \mathbb{1}_{[\epsilon_{i,r,A}, +\infty)}(u).$$

Define $\min \in \{1, 2\}$ such that $\epsilon_{\min} = \epsilon_{\min,r,A} = \min\{\epsilon_{1,r,A}, \epsilon_{2,r,A}\}$ and similarly $\max \in \{1, 2\}$. Then

$$\mathcal{W}(\rho_{1,r,A}, \rho_{2,r,A}) = \int_{\epsilon_{\min}}^{\epsilon_{\max}} F_{\min,r,A}(u)\, du + \int_{\epsilon_{\max}}^{+\infty} |F_{1,r,A}(u) - F_{2,r,A}(u)|\, du.$$

Now, $r \int_{\epsilon_{\min}}^{\epsilon_{\max}} F_{\min,r,A}(u)\, du = \int_{\epsilon_{\min}}^{\epsilon_{\max}} \nu_{\min}([\epsilon_{\min}, u) \times A)\, dy \leq (\epsilon_{\max} - \epsilon_{\min})\, r$, which can be rewritten as $\epsilon_{\max}\nu_{\max,x}([\epsilon_{\max}, +\infty)) - \epsilon_{\min}\nu_{\min,x}([\epsilon_{\min}, +\infty))$. Thus by Lemma 18 it converges to zero as r goes to $+\infty$. On the other hand,

$$r \int_{\epsilon_{\max}}^{+\infty} |F_{1,r,A}(u) - F_{2,r,A}(u)|\, du = \int_{\epsilon_{\max}}^{+\infty} |\nu_1((u, +\infty) \times A) - \nu_2((u, +\infty) \times A)|\, du,$$

which attains the expression for $g_u(A)$ as r goes to $+\infty$.

2.6.2 Proof of Corollary 6

For every $u \in \mathbb{R}^+$, $\nu_h((u, +\infty) \times A) \leq \nu_j((u, +\infty) \times A)$ because the Radon–Nikodym derivatives are ordered. Thus $g_u(A)$ is equal to

$$\begin{aligned} &\left| \int_0^{+\infty} (\nu_1((u, +\infty) \times A) - \nu_2((u, +\infty) \times A))\, du \right| \\ &= \left| \int_0^{+\infty} \int_u^{+\infty} (\nu_1(ds \times A) - \nu_2(ds \times A))\, du \right|. \end{aligned}$$

By interchanging the integrals this is equal to the lower bound in Theorem 5.

2.6.3 Proof of Corollary 7

Without loss of generality we assume $\alpha_1(A) \leq \alpha_2(A)$. Then by taking $\eta(ds) = \mathbb{1}_{(0,+\infty)}(s)\, ds$,

$$\nu_1(s \times A) = \alpha_1(A)\, \rho(s) \leq \alpha_2(A)\, \rho(s) = \nu_2(s \times A),$$

for every $s \in \mathbb{R}^+$. Thus $\mathcal{W}(\tilde{\mu}_1(A), \tilde{\mu}_2(A)) = |\alpha_1(A) - \alpha_2(A)| \int_{\mathbb{R}^+} s\, \rho(s)\, ds$ by Corollary 6. We conclude by taking the supremum over $A \in \mathcal{X}$.

2.6.4 Proof of Proposition 8

Let $\tilde{\mu}_i \sim \mathrm{Ga}(b_i, \alpha_i)$, for $i = 1, 2$. Without loss of generality we assume $0 < b_1 \leq b_2$. Thus for every $A \in \mathcal{X}$,

$$\mathbb{E}(\tilde{\mu}_i(A)) = \int_{\mathbb{R}^+ \times A} s\, \nu_i(ds, dy) = \frac{\alpha_i(A)}{b_i}.$$

This implies that the total mean is finite. Since $\frac{e^{-sb_i}}{s}$ is not integrable near zero, the random measures are infinitely active. Thus Theorem 5 holds and from the expression of $\mathbb{E}(\tilde{\mu}_i(A))$ above we derive the lower bound in Proposition 8. We now focus on the upper bound. For every $u \in \mathbb{R}^+$,

$$\nu_i((u, +\infty) \times A) = \alpha_i(A) \int_u^{+\infty} \frac{e^{-sb_i}}{s} ds = \alpha_i(A)\, E_1(b_i u),$$

where $E_1(x) = \int_x^\infty \frac{e^{-y}}{y} dy$. Thus,

$$\mathcal{W}(\tilde{\mu}_1(A), \tilde{\mu}_2(A) \leq \int_0^{+\infty} |\alpha_1(A)\, E_1(b_1 u) - \alpha_2(A)\, E_1(b_2 u)|\, du.$$

The fundamental theorem of line integral ensures that

$$\int_0^{+\infty} |\alpha_1(A)\, E_1(b_1 u) - \alpha_2(A)\, E_1(b_2 u)|\, du = \int_0^{+\infty} \Big| \int_C \nabla \psi_u(a, b) \cdot ds \Big| du,$$

where ∇ denotes the gradient of a function, $\psi_u(a, b) = a\, E_1(by)$, and C is the segment in $\mathbb{R}^2$ connecting $(\alpha_1(A), b_1)$ to $(\alpha_2(A), b_2)$. We consider the parametrization $s(t) = (\alpha_1(A) + t(\alpha_2(A) - \alpha_1(A)), b_1 + t(b_2 - b_1))$. Since $\nabla \psi_u(a, b) = (E_1(by), -\frac{a}{b} e^{-by})$, this is equal to

$$\begin{aligned}
&\int_0^{+\infty} \Big| \int_0^1 \Big(E_1(s_2(t)u)\, s_1'(t) - \frac{s_1(t)}{s_2(t)} e^{-s_2(t)u}\, s_2'(t) \Big)\, dt \Big|\, du \\
&\leq \int_0^{+\infty} \int_0^1 \Big| E_1((b_1 + t(b_2 - b_1))u)\, (\alpha_2(A) - \alpha_1(A)) - \\
&\qquad\qquad + \frac{\alpha_1(A) + t(\alpha_2(A) - \alpha_1(A))}{b_1 + t(b_2 - b_1)}\, e^{-(b_1 + t(b_2 - b_1))u}\, (b_2 - b_1) \Big| dt\, du
\end{aligned}$$

Since we have assumed w.l.o.g. that $b_1 \leq b_2$,

$$\begin{aligned}
&\leq \int_0^{+\infty} \int_0^1 \Big(E_1((b_1 + t(b_2 - b_1))u)\, |\alpha_2(A) - \alpha_1(A)| + \\
&\qquad\qquad + \frac{\alpha_1(A) + t(\alpha_2(A) - \alpha_1(A))}{b_1 + t(b_2 - b_1)} e^{-(b_1 + t(b_2 - b_1))u}\, (b_2 - b_1) \Big) dt\, du
\end{aligned}$$

We invert the integrals thanks to Fubini's Theorem and use the fact that $\int_0^{+\infty} E_1(ax)dx = \frac{1}{a}$, which is a standard result on exponential integrals (Geller & W. Ng, 1969). Thus,

$$\leq \int_0^1 \Big(\frac{1}{b_1 + t(b_2 - b_1)} \left|\alpha_2(A) - \alpha_1(A)\right| + \frac{\alpha_1(A) + t(\alpha_2(A) - \alpha_1(A))}{(b_1 + t(b_2 - b_1))^2} (b_2 - b_1) \Big) dt.$$

By standard integration techniques, this amounts to

$$= \frac{\alpha_1(A)}{b_1} - \frac{\alpha_2(A)}{b_2} + \mathbb{1}_{(0,+\infty)}(\alpha_2(A) - \alpha_1(A))\, 2\, \frac{\alpha_2(A) - \alpha_1(A)}{b_2 - b_1} \log \frac{b_2}{b_1}.$$

2.6.5 Proof of Proposition 9

When $c_1 = c_2$ the result is immediate once we observe that $\mathbb{E}(\tilde{\mu}(A)) = \alpha(A)$ for every $\tilde{\mu} \sim \mathrm{Be}(c, \alpha)$. We focus on the case $\alpha_1 = \alpha_2$, $0 < c_1 \leq c_2$. By reasoning as in the proof of Proposition 8, $\mathcal{W}(\tilde{\mu}_1(A), \tilde{\mu}_2(A)$ is bounded from above by

$$\alpha(A) \int_0^1 \int_0^s \int_{c_1}^{c_2} \left| \frac{d}{dc} \left(\frac{c\,(1-s)^{c-1}}{s} \right) \right| dc\, du\, ds$$

the derivative $\left((c\,(1-s)^{c-1})\, s^{-1}\right)' = (1-s)^{c-1}\,(1 + c\,\log(1-s))s^{-1} \leq (1-s)^{c-1}\,(1 - c\,\log(1-s))s^{-1}$ for $s \in (0,1)$. Thus by Fubini's Theorem the previous integral is bounded from above by

$$\alpha(A) \int_{c_1}^{c_2} \int_0^1 (1-s)^{c-1}\,(1 - c\,\log(1-s))\, ds\, dc = 2\alpha(A) \log\left(\frac{c_2}{c_1}\right),$$

by standard integration techniques.

2.6.6 Proof of Lemma 10

Let $\{A_1, \cdots A_n\}$ in $\mathcal{X}$ be disjoint sets. Then for $i = 1, \ldots, n$ the random variables $\tilde{\mu}_f(A_i) = \int_{A_i} f(x)\tilde{\mu}(dx)$ are independent since f is deterministic and $\tilde{\mu}(A_1), \cdots \tilde{\mu}(A_n)$ are independent. This proves that $\tilde{\mu}_f$ is a CRM. In order to find its Lévy intensity ν_f, we consider the Laplace functional transform (1.5):

$$\begin{aligned}\mathbb{E}\big(e^{-\int_{\mathbb{R}} g(y)\tilde{\mu}_f(dy)}\big) &= \exp\Big\{ -\int_{\mathbb{R}^+\times\mathbb{R}} [1 - e^{-s\,g(y)f(y)}]\,\nu(ds, dy)\Big\} = \\ &= \exp\Big\{ -\int_{\mathbb{R}^+\times\mathbb{R}} [1 - e^{-s\,g(y)}]\,(p_f \,\#\, \nu)(ds, dy)\},\end{aligned}$$

where $p_f(s, y) = (sf(y), y)$.

2.6.7 Proof of Theorem 13

The lower bound follows immediately from (2.3). As for the upper bounds, first of all we observe that for any $x, y \in \mathbb{R}^+$

$$|e^{-x} - e^{-y}| = e^{-x\wedge y}(1 - e^{-|x-y|}). \tag{2.27}$$

In order to derive the function $g_{u,1}$ in the upper bound, we observe that since $e^{-x\wedge y} \leq 1$, the following upper bound holds for $\mathcal{W}(\tilde{S}_1(t), \tilde{S}_2(t))$:

$$\inf_{C(\tilde{H}_1(t),\tilde{H}_2(t))} \mathbb{E}\big(\big|e^{-\tilde{H}_1(t)} - e^{-\tilde{H}_2(t)}\big|\big) \leq \inf_{C(\tilde{H}_1(t),\tilde{H}_2(t))} \mathbb{E}\big(1 - e^{-|\tilde{H}_1(t)-\tilde{H}_2(t)|}\big).$$

Since $1 - e^{-x}$ is a concave function, by Jensen's inequality, $\mathcal{W}(\tilde{S}_1(t), \tilde{S}_2(t))$ is bounded from above by

$$\inf_{C(\tilde{H}_1(t),\tilde{H}_2(t))} \Big\{1 - e^{-\mathbb{E}(|\tilde{H}_1(t)-\tilde{H}_2(t)|)}\Big\} = 1 - e^{-\inf_{C(\tilde{H}_1(t),\tilde{H}_2(t))}\mathbb{E}(|\tilde{H}_1(t)-\tilde{H}_2(t)|)}.$$

As for $g_{u,2}$, combining (2.27) with $e^{-x\wedge y} \leq e^{-x} + e^{-y}$ and by Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}\big(\big|e^{-\tilde{H}_1(t)} - e^{-\tilde{H}_2(t)}\big|\big) \\
&\leq \mathbb{E}\big(e^{-\tilde{H}_1(t)} + e^{-\tilde{H}_2(t)} - e^{-(\tilde{H}_1(t)+|\tilde{H}_1(t)-\tilde{H}_2(t)|)} - e^{-(\tilde{H}_2(t)+|\tilde{H}_1(t)-\tilde{H}_2(t)|)}\big) \\
&\leq \mathbb{E}\big(e^{-\tilde{H}_1(t)}\big) + \mathbb{E}\big(e^{-\tilde{H}_2(t)}\big) - e^{-\mathbb{E}(\tilde{H}_1(t)+|\tilde{H}_1(t)-\tilde{H}_2(t)|)} - e^{-\mathbb{E}(\tilde{H}_2(t)+|\tilde{H}_1(t)-\tilde{H}_2(t)|)} \\
&\leq \mathbb{E}\big(e^{-\tilde{H}_1(t)}\big) + \mathbb{E}\big(e^{-\tilde{H}_2(t)}\big) - \big(e^{-\mathbb{E}(\tilde{H}_1(t))} + e^{-\mathbb{E}(\tilde{H}_2(t))}\big)e^{-\mathbb{E}|\tilde{H}_1(t)-\tilde{H}_2(t)|}.
\end{aligned}$$

By taking the infimum over all couplings in $C(\tilde{H}_1(t), \tilde{H}_2(t))$ we derive $g_{u,2}$.

2.6.8 Proof of Theorem 15

The proof relies on $\gamma(y)$ being a decreasing continuous piecewise linear function. We first include $x_1, \ldots, x_n$ in the set $\{t_h \,|\, h = 1, \ldots, n(\epsilon)\}$. Then, for every $i = 1, \ldots n$ we iteratively include all points $t \in (x_i, x_{i+1})$ such that the counterimage $\gamma^{-1}(t)$ is at distance ϵ from the previous point. We easily conclude by observing that on $(x_i, x_{i+1}]$ the function γ is linear with coefficient equal to $-(n - i)$. See Figure 2.3.

2.6.9 Proof of Theorem 17

The proof is based on observing that $\tilde{h}_1 = \tilde{h}^*$ and $\tilde{h}_2 = \tilde{h}^\epsilon$ are two kernel mixture hazards with $k_1(y|t) = k_2(y|t) = \mathbb{1}_{[0,t]}(y)$ and $\tilde{\mu}_i$ extended gamma CRMs with scale function $\beta_1(y) = \beta^*(y)$ and $\beta_2(y) = \beta^\epsilon(y)$ and Lebesgue base measure on the positive axis, i.e.

$$\nu_i(ds, dy) = \frac{\exp\big\{-\frac{s}{\beta_i(y)}\big\}}{s}\,\mathbb{1}_{[0,+\infty)}(y)\,ds\,dy,$$

These kernel and Lévy measures satisfy both the conditions of Theorem 11 and of Theorem 12. Since by construction $\beta^\epsilon(y) \leq \beta^*(y)$ for every $y \in [0, +\infty)$, the Lévy densities

of the the hazards and of the cumulative hazards are ordered. Thus the Wasserstein distance reduces to the absolute difference of their means:

$$\mathcal{W}(\tilde{h}^*(t), \tilde{h}^\epsilon(t)) = \left| \int_0^t \beta^*(y) - \beta^\epsilon(y)\, dy \right| \leq \int_0^t |\beta^*(y) - \beta^\epsilon(y)|\, dy \leq \epsilon \frac{\beta^2}{b^2}\, t,$$

by (2.24). Similarly,

$$\mathcal{W}(\tilde{H}^*(t), \tilde{H}^\epsilon(t)) = \left| \int_0^t (t-y)(\beta^*(y) - \beta^\epsilon(y))\, dy \right| \leq \epsilon \frac{\beta^2}{b^2} \int_0^t (t-y) dy = \epsilon \frac{\beta^2}{2b^2}\, t^2.$$

Finally, the bound for the survival function derives directly from the one on the cumulative hazards, as in Theorem 13.

Chapter 3

Measuring dependence in the Wasserstein distance

The proposal and study of dependent Bayesian nonparametric models has been one of the most active research lines in the last two decades, with random vectors of measures representing a natural and popular tool to define them. Nonetheless a principled approach to understand and quantify the associated dependence structure is still missing. In this chapter we devise a general, and non model-specific, framework to achieve this task for random measure based models, which consists in: (a) quantify dependence of a random vector of probabilities in terms of closeness to exchangeability, which corresponds to the maximally dependent coupling with the same marginal distributions, i.e. the comonotonic vector; (b) recast the problem in terms of the underlying random measures (in the same Fréchet class) and quantify the closeness to comonotonicity; (c) define a distance based on the Wasserstein metric, which is ideally suited for spaces of measures, to measure the dependence in a principled way. Several results, which represent the very first in the area, are obtained. In particular, useful bounds in terms of the underlying Lévy intensities are derived relying on compound Poisson approximations. These are then specialized to popular models in the Bayesian literature leading to interesting insights.

3.1 Introduction

A sequence of random elements $(X_n)_{n\geq 1}$ is exchangeable when its distribution is invariant with respect to finite permutations of the indices. By de Finetti's Representation Theorem this intuitive symmetry requirement is equivalent to the finite–dimensional distributions being conditionally independent and identically distributed. Partial exchangeability (de Finetti, 1938) is a natural generalization and corresponds to exchangeability holding within each of a finite number of blocks in which the random elements are grouped. The corresponding representation theorem states that for partially exchangeable sequences $\{X_{1,j} \,|\, j \geq 1\}, \ldots, \{X_{k,j} \,|\, j \geq 1\}$ on a Polish space $\mathbb{X}$ there exists a random vector of probability measures $(\tilde{p}_1, \ldots, \tilde{p}_k) \sim Q$ s.t. for any $n_i \in \mathbb{N}$ and any

Borel sets $A_i \subset \mathbb{X}^{n_i}$, for $i = 1, \dots, k$,

$$\mathbb{P}\Big(\bigcap_{i=1}^{k} \{(X_{i,1}, \cdots, X_{i,n_i}) \in A_i\} \Big) = \int_{P_{\mathbb{X}}^k} \prod_{i=1}^{k} p_i^{(n_i)}(A_i)\, Q(dp_1, \dots, dp_k).$$

In particular, exchangeability is recovered when $\tilde{p}_1 = \cdots = \tilde{p}_k$ almost surely (a.s.). We refer to Section 1.5 for further details.

In Bayesian nonparametric inference, the random elements $\{X_{1,j} \,|\, j \geq 1\}, \dots, \{X_{k,j} \,|\, j \geq 1\}$ are regarded as observables and a fundamental issue is the choice of the distribution Q for the random vector of probability measures $(\tilde{p}_1, \dots, \tilde{p}_k)$, the prior distribution. The dependence between the random probabilities is of crucial importance, since it regulates the dependence between groups of observations and, consequently, the borrowing of information across groups. The first proposal of a dependent nonparametric prior dates back to Cifarelli & Regazzini (1978), but it were the two seminal papers of MacEachern (1999, 2000) which led to an impressive growth of research in this direction. Most classes of priors are defined to select a.s. discrete $\tilde{p}_i$'s, since this naturally allows for clustering at either the observations' or latent level. This is true also in the exchangeable case: an a.s. discrete $\tilde{p}$ is obtained through either the stick–breaking construction (Sethuraman, 1994; Ishwaran & James, 2001) or a suitable transformation of a completely random measure (CRM) $\tilde{\mu}$ (Kingman, 1967; Lijoi & Prünster, 2010). The former approach is particularly effective for computational purposes, whereas the latter allows to derive important distributional properties. In particular, by using CRMs as a unifying concept, as showcased in (Lijoi & Prünster, 2010), one obtains popular classes of nonparametric priors such as, e.g., normalized random measures (Regazzini et al., 2003), neutral-to-the-right processes (Doksum, 1974) and kernel mixtures of random measures (Dykstra & Laud, 1981; James, 2005). Correspondingly, in the general partially exchangeable case, one may distinguish two approaches for building dependent priors: the first approach models the dependence at the level of the atoms and/or the jumps of the stick–breaking construction of each $\tilde{p}_i$; the second models the dependence at the level of the CRMs $(\tilde{\mu}_1, \dots, \tilde{\mu}_k)$ to then obtain a dependent vector $(\tilde{p}_1, \dots, \tilde{p}_k)$ via a suitable transformation. See (Hjort et al., 2010; Müller et al., 2015; Ghosal & van der Vaart, 2017; Müller et al., 2018) for extensive accounts.

A crucial gap in this vast literature is the understanding and quantification of the dependence structure of a dependent nonparametric prior in order to both elicit prior parameters to achieve the desired degree of dependence and compare different priors themselves. The most natural way to approach the problem is to measure closeness to exchangeability, which corresponds to the extreme case of maximal dependence between populations. Within a parametric framework, already in 1938, de Finetti proposed to use *approximately* exchangeable priors to deal with contingency tables (de Finetti, 1938). Recently, Bacallado et al. (2015) enriched this class of examples and proposed ways to use them to test for the exchangeability assumption. However, closeness to exchangeability is left as an essentially intuitive notion. To the best of our knowledge, the only measure of dependence that has been used so far is the pairwise linear correlation of $(\tilde{p}_i(A), \tilde{p}_j(A))$, for any given set A, which is certainly useful but reducing dependencies

between random probabilities to linear correlation is hardly satisfying.
Here, we tackle the problem in a general nonparametric framework adopting a principled approach in that we measure the distance to exchangeability in terms of the Wasserstein distance. Because of its intrinsically geometric definition, the Wasserstein distance is the most appropriate choice for describing the similarity between distributions. As explained in (Rachev, 1985), this distance was first introduced by Gini (1914) with this exact scope. During the past century the Wasserstein distance was introduced and studied in many fields of research, including Optimal Transport Theory, Partial Differential Equations and Ergodic Theory. Recently, it has gained a renewed popularity in Probability, Statistics and the related fields of Machine Learning and Optimization, where the distinguished theoretical properties are now supported by efficient algorithms (Cuturi, 2013). See (Villani, 2008; Panaretos & Zemel, 2019) for detailed reviews. The first to use the Wasserstein distance in a Bayesian nonparametric framework, for asymptotic investigations, has been Nguyen (2013) who has convincingly argued for it as an effective tool to handle discrete nonparametric priors. See also (Nguyen, 2016). From our perspective, the Wasserstein distance is the ideal choice because it allows for a meaningful comparison between distributions with different support and without density, as the ones arising from transformations of CRMs. This property is not shared by the most common distances and divergences, such as the total variation distance, the Hellinger distance or the Kullback–Leibler divergence.
Our general setup is as follows. For simplicity, we consider the case $k = 2$, even though most of our results may be extended to a generic k with no additional cost. Since our leading purpose is to measure the closeness to exchangeability (i.e. $\tilde{p}_1 = \tilde{p}_2$ a.s), we consider random vectors $(\tilde{p}_1, \tilde{p}_2)$ with equal marginal distributions ($\tilde{p}_1 \overset{\mathrm{d}}{=} \tilde{p}_2$). A crucial observation is then the following: instead of measuring the distance from exchangeability of $(\tilde{p}_1, \tilde{p}_2)$, we work with *completely random vectors* (CRVs) $(\tilde{\mu}_1, \tilde{\mu}_2)$, characterized by jointly independent increments, and measure their closeness to the comonotonic case i.e. $\tilde{\mu}_1 = \tilde{\mu}_2$ a.s. In fact, since most random vectors of discrete probabilities $(\tilde{p}_1, \tilde{p}_2)$ are obtained by a suitable component–wise transformation T of a CRV, $(T(\tilde{\mu}_1), T(\tilde{\mu}_2))$ (see (Lijoi & Prünster, 2010) for details), comonotonic CRVs correspond to exchangeability. Working directly with the random measures rather than their transformed versions has two distinct advantages: (a) it provides a generic and non-model specific framework for the analysis of dependence, which can then be tailored to the particular class of models one is interested in, as we do in Section 3.7; (b) it significantly simplifies the mathematical analysis. Closeness to the comonotonic case is then measured through the following distance on CRVs, which will be shown in Section 3.2 to be well–defined,

$$d_{\mathcal{W}}\left(\begin{pmatrix}\tilde{\mu}_1\\ \tilde{\mu}_2\end{pmatrix}, \begin{pmatrix}\tilde{\xi}_1\\ \tilde{\xi}_2\end{pmatrix}\right) = \sup_{A\in\mathcal{X}} \mathcal{W}\left(\begin{pmatrix}\tilde{\mu}_1(A)\\ \tilde{\mu}_2(A)\end{pmatrix}, \begin{pmatrix}\tilde{\xi}_1(A)\\ \tilde{\xi}_2(A)\end{pmatrix}\right), \tag{3.1}$$

where $\mathcal{W}$ denotes the 2–Wasserstein distance on the Euclidean plane. The main goal of this work is then to provide an analytical expression for the distance $d_{\mathcal{W}}$ in (3.1) with a particular focus on the distance between a CRV $(\tilde{\mu}_1, \tilde{\mu}_2)$ and the comonotonic random vector $(\tilde{\xi}_1, \tilde{\xi}_2)$ in the same Fréchet class, i.e. with the same marginal distributions. We

stress that our results, even though motivated by Bayesian nonparametric models, are of independent probabilistic interest with reference to the theory of multidimensional random measures and Lévy processes.
The two major challenges in the treatment of $d_{\mathcal{W}}$ in (3.1) may be summarized as follows. (i) The analytical computation of the Wasserstein distance needs the appointment of an optimal transport map. While these are always known in explicit form for univariate distributions, the general expression for multidimensional ones is still an open problem, with only a few known cases. Knott & Smith (1984) characterized optimal mappings as gradients of convex functions. By relying on a reformulation of this result by Rüschendorf (1991), in Theorem 20 we find the optimal transport map to the comonotonic vector. This allows to express the Wasserstein distance as an integral that requires the cumulative distribution function (cdf) of $\tilde{\mu}_1(A) + \tilde{\mu}_2(A)$, which in some cases may be computed directly, as for the Wasserstein distance between comonotonicity and independence. (ii) The law of a CRV is usually provided through a bivariate Lévy measure, so that the cdf of $\tilde{\mu}_1(A) + \tilde{\mu}_2(A)$ is not available in closed form. In Theorem 23 we find tight bounds of the distance that are expressed directly in terms of the Lévy measures. This is achieved through suitable compound Poisson approximations of the random vectors and by finding a new informative bound for the Wassserstein distance between multivariate compound Poisson distributions (Proposition 24). Much effort is then put in the computation of the bounds for $d_{\mathcal{W}}$ when $(\tilde{\mu}_1, \tilde{\mu}_2)$ is taken to be equal to well–known priors in Bayesian nonparametric models for partially exchangeable data, leading to meaningful insights and a quantification of their dependence structure in terms of hyperparameters.
Our measure of dependence may be naturally extended to $k > 2$ groups by considering the Wasserstein distance on $\mathbb{R}^k$ from $(\xi_1(A), \dots, \xi_k(A))$ such that $\xi_1(A) = \cdots = \xi_k(A)$ a.s., i.e. the comonotonic k-dimensional CRV corresponding to exchangeability. The main techniques described in the previous paragraph continue to hold in the k–dimensional case and for simplicity we focus on $k = 2$. We underline that the natural extension to an arbitrary k provides a further benefit of our measure of dependence compared to linear correlation, since it provides an overall quantification of dependence without forcing pairwise comparisons.
Many of the techniques that we introduce may also be used to measure the dependence directly on component-wise tranformations $(T(\tilde{\mu}_1), T(\tilde{\mu}_2))$ of a CRV. However, this requires additional work and depends on the choice of T, since the Wasserstein distance in not invariant with respect to transformations. We develop informative bounds for a specific transformation that is widely used in Bayesian nonparametric models for time-to-event data, where random hazards are often modeled as kernel mixtures over a CRM. Since the hazards characterize the entire distribution, this provides a specification for the de Finetti measure. The inferential properties of this prior were thouroughly studied by Dykstra & Laud (1981); Lo & Weng (1989) and James (2005) for exchangeable observations and have seen interesting generalizations to a partially exchangeable setting (Lijoi & Nipoti, 2014; Camerlenghi et al., 2020).

The chapter is structured as follows. In Section 3.2 we introduce necessary concepts

and notation and prove that $d_{\mathcal{W}}$ is actually a distance. In Section 3.3 we obtain an integral representation of the Wasserstein distance between a random vector of measures and the corresponding comonotonic one. In Section 3.4 we develop general bounds for the distance between CRVs in the same Fréchet class, in terms of their bivariate Lévy intensities. In Section 3.5 we focus on the distance from exchangeability and obtain an explicit form for the bounds of the previous section. In particular, in Section 3.6 we use them to bound the distance between exchangeability and the other extreme case, independence. Finally, in Section 3.7 the previous techniques are used to quantify the dependence of three popular nonparametric priors for partially exchangeable data, namely compound random measures (Griffin & Leisen, 2017; Riva Palacio & Leisen, 2019), Clayton Lévy copula (Tankov, 2003; Epifani & Lijoi, 2010; Leisen & Lijoi, 2011) and GM–dependence (Griffiths & Milne, 1978; Lijoi et al., 2014; Lijoi & Nipoti, 2014). In Section 3.8 we extend the measure of dependence to random hazards that are modeled as kernel mixtures over a CRV, with a specific application to GM-dependence (Lijoi & Nipoti, 2014) All proofs are deferred to Section 3.9.

3.2 Preliminaries

We first recall definitions and key properties of random vectors of measures and of the Wasserstein distance. To fix notation, let $\mathbb{R}_+ = (0, +\infty)$ and $\mathbb{R}_+^2 := [0, +\infty) \times [0, +\infty) \setminus \{(0, 0)\}$. Moreover, $\mathcal{L}(X)$ denotes the law of a random variable X and $\overset{\mathrm{d}}{=}$ refers to equality in distribution.
Let $(\mathbb{X}, d_{\mathbb{X}})$ be a Polish space endowed with a distance $d_{\mathbb{X}}$ and the Borel σ-algebra $\mathcal{X}$. We denote by $(M_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$ the Borel space of boundedly finite measures on $\mathbb{X}$ endowed with the topology of weak$^\sharp$ convergence (Daley & Vere-Jones, 2002). A random vector of measures is a measurable function $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \tilde{\mu}_2) : \Omega \to M_{\mathbb{X}}^2$, where $(\Omega, \Sigma_\Omega, P_\Omega)$ is a generic probability space and $M_{\mathbb{X}}^2 = M_{\mathbb{X}} \times M_{\mathbb{X}}$ is endowed with the product σ–algebra. We refer to the projections $\pi_i \circ \tilde{\boldsymbol{\mu}} = \tilde{\mu}_i : \Omega \to M_{\mathbb{X}}$, for $i = 1, 2$, as the marginals of $\tilde{\boldsymbol{\mu}}$. Moreover, the random vectors of one–dimensional distributions are denoted as $\tilde{\boldsymbol{\mu}}(A) = (\tilde{\mu}_1(A), \tilde{\mu}_2(A)) : \Omega \to [0, +\infty) \times [0, +\infty)$, for every $A \in \mathcal{X}$.

Definition 6. A random vector of measures $\tilde{\boldsymbol{\mu}}$ is a *completely random vector* (CRV), if given a finite collection of disjoint bounded Borel sets $\{A_1, \cdots, A_n\}$, the random vectors $\{\tilde{\boldsymbol{\mu}}(A_1), \ldots, \tilde{\boldsymbol{\mu}}(A_n)\}$ are independent.

In particular, this definition entails that the marginal distributions $\tilde{\mu}_1, \tilde{\mu}_2$ have independent increments and are thus completely random measures (CRMs) in the sense of Kingman (1967), as introduced in Section 1.4. We point out that the converse is not necessarily true: a random vector of measures whose marginals are CRMs is not necessarily a CRV. The joint independence of the increments guarantees that the distribution of $\tilde{\boldsymbol{\mu}}$ is characterized by the distribution of the one–dimensional random vectors $\{\tilde{\boldsymbol{\mu}}(A) \,|\, A \in \mathcal{X}\}$. Moreover, (Kallenberg, 2017, Theorem 3.19) ensures that, if $\tilde{\boldsymbol{\mu}}$ has no

fixed atoms, there exists a Poisson random measure $\mathcal{N}$ on $\mathbb{R}_+^2 \times \mathbb{X}$ s.t. for every $A \in \mathcal{X}$,

$$\tilde{\boldsymbol{\mu}}(A) \stackrel{\mathrm{d}}{=} \int_{\mathbb{R}_+^2 \times A} \boldsymbol{s}\, \mathcal{N}(ds_1, ds_2, dx), \tag{3.2}$$

where $\boldsymbol{s} = (s_1, s_2)$. The mean measure $\nu(ds_1, ds_2, dx) = \mathbb{E}(\mathcal{N}(ds_1, ds_2, dx))$ satisfies the following properties: $\nu(\mathbb{R}_+^2 \times \{x\}) = 0$ for every $x \in \mathbb{X}$ and

$$\int_{\mathbb{R}_+^2 \times A} \min\{s_1 + s_2, \epsilon\}\, \nu(ds_1, ds_2, dx) < +\infty \tag{3.3}$$

for every bounded $A \in \mathcal{X}$ and every $\epsilon > 0$. We will focus on CRVs without fixed atoms and refer to ν as the *intensity measure* of $\tilde{\boldsymbol{\mu}}$. This will be further assumed to have no atoms. Campbell's Theorem ensures that from the Lévy intensity of $\tilde{\boldsymbol{\mu}}$ one derives the Lévy intensities of the marginal CRMs $\tilde{\mu}_1$ and $\tilde{\mu}_2$, namely

$$\nu_1(ds, dx) = \int_{[0,+\infty)} \nu(ds, ds_2, dx), \qquad \nu_2(ds, dx) = \int_{[0,+\infty)} \nu(ds_1, ds, dx).$$

We underline that the marginal CRMs are not forced to have the same atoms a.s. because the measure ν may have positive mass on the axes, as it will be clear from Section 3.6. We say that $\tilde{\boldsymbol{\mu}}$ is *infinitely active* if for every $A \in \mathcal{X}$ both the marginal CRMs are infinitely active, i.e.

$$\int_{\mathbb{R}_+ \times A} \nu_1(ds, dx) = \int_{\mathbb{R}_+ \times A} \nu_2(ds, dx) = +\infty. \tag{3.4}$$

Since most applications of random measures in Bayesian nonparametrics deal with infinitely active random measures, we concentrate on these.

The distribution of a CRV is characterized by the distribution of the one–dimensional distributions $\{\tilde{\boldsymbol{\mu}}(A) \,|\, A \in \mathcal{X}\}$. Thus any distance $\mathcal{D}$ on the space $\mathcal{P}_{\mathbb{R}^2}$ of probability measures on $\mathbb{R}^2$ determines a distance on the laws of CRVs by considering

$$\sup_{A \in \mathcal{X}} \mathcal{D}(\mathcal{L}(\tilde{\boldsymbol{\mu}}^1(A)), \mathcal{L}(\tilde{\boldsymbol{\mu}}^2(A))).$$

The distance $d_{\mathcal{W}}$ defined in (3.1) fits in this general framework, by considering the Wasserstein distance as metric $\mathcal{D}$. Given π_1, π_2 two probability measures on a Polish space $(\mathbb{X}, d_{\mathbb{X}})$, we indicate by $C(\pi_1, \pi_2)$ the Fréchet class of π_1 and π_2, i.e. the set of distributions on the product space whose marginal distributions coincide with π_1 and π_2 respectively. If Z_1 and Z_2 are dependent random variables on $\mathbb{X}$ such that their joint law $\mathcal{L}(Z_1, Z_2) \in C(\pi_1, \pi_2)$, we write $(Z_1, Z_2) \in C(\pi_1, \pi_2)$.

Definition 7. The Wasserstein distance of order $p \in [1, +\infty)$ between π_1 and π_2 is

$$\mathcal{W}_{p, d_{\mathbb{X}}}(\pi_1, \pi_2) = \inf_{(Z_1, Z_2) \in C(\pi_1, \pi_2)} \left\{ \mathbb{E}(d_{\mathbb{X}}(Z_1, Z_2)^p) \right\}^{\frac{1}{p}}.$$

By extension, we refer to the Wasserstein distance between two random elements $X_i : \Omega \to \mathbb{X}$, $i = 1, 2$, as the Wasserstein distance between their laws, i.e. $\mathcal{W}_{p,d}(X_1, X_2) = \mathcal{W}_{p,d}(\mathcal{L}(X_1), \mathcal{L}(X_2))$. An element of $C(\mathcal{L}(X_1), \mathcal{L}(X_2))$ is referred to as a coupling between X_1 and X_2.

Throughout the work we set $p = 2$ and $(\mathbb{X}, d_{\mathbb{X}}) = (\mathbb{R}^2, \|\cdot\|)$, i.e. the Euclidean plane. We will refer to such distance as the Wasserstein distance and denote it by $\mathcal{W}$, i.e.

$$\mathcal{W}(\boldsymbol{X}, \boldsymbol{Y}) = \inf_{(\boldsymbol{Z}_X, \boldsymbol{Z}_Y) \in C(\boldsymbol{X}, \boldsymbol{Y})} \left\{ \mathbb{E}(\|\boldsymbol{Z}_X - \boldsymbol{Z}_Y\|^2) \right\}^{\frac{1}{2}},$$

where we have used the vector notation $\boldsymbol{X} = (X_1, X_2) \in \mathbb{R}^2$. The parallelogram rule on normed spaces ensures that

$$\mathcal{W}(\boldsymbol{X}, \boldsymbol{Y})^2 \leq 2 \left(\mathbb{E}(\|\boldsymbol{X}\|^2) + \mathbb{E}(\|\boldsymbol{Y}\|^2) \right). \tag{3.5}$$

In particular, the Wasserstein distance between random elements on $\mathbb{R}^2$ with finite expected squared norm is finite. Thus, in order for $d_{\mathcal{W}}$ in (3.1) to be finite, we restrict to random vectors of measures with finite second moment $\mathbb{E}(\|\tilde{\boldsymbol{\mu}}(\mathbb{X})\|^2) = \mathbb{E}(\tilde{\mu}_1(\mathbb{X})^2) + \mathbb{E}(\tilde{\mu}_2(\mathbb{X})^2) < +\infty$. Therefore, by standard properties of Poisson random measures, we ask

$$\mathbb{E}(\tilde{\boldsymbol{\mu}}(\mathbb{X})) = \int_{\mathbb{R}_+^2 \times \mathbb{X}} \boldsymbol{s}\, \nu(ds_1, ds_2, dx) < +\boldsymbol{\infty}, \tag{3.6}$$

$$\mathrm{Var}(\tilde{\boldsymbol{\mu}}(\mathbb{X})) = \int_{\mathbb{R}_+^2 \times \mathbb{X}} \boldsymbol{s}^2\, \nu(ds_1, ds_2, dx) < +\boldsymbol{\infty}, \tag{3.7}$$

where $\boldsymbol{s}^2 = (s_1^2, s_2^2)$ and $+\boldsymbol{\infty} = (+\infty, +\infty)$. We summarize our findings in the following.

Proposition 19. *The function $d_{\mathcal{W}} : \mathbb{P}(M_{\mathbb{X}}^2) \times \mathbb{P}(M_{\mathbb{X}}^2) \to [0, +\infty)$ defines a distance on the laws of CRVs whose Lévy intensities satisfy (3.6) and (3.7).*

We conclude this section by recalling some properties of the Wasserstein distance to be used in the sequel. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random elements in $\mathbb{R}^2$. A coupling $(\boldsymbol{Z}_X, \boldsymbol{Z}_Y) \in C(\boldsymbol{X}, \boldsymbol{Y})$ is said to be optimal if $\mathcal{W}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}(\|\boldsymbol{Z}_X - \boldsymbol{Z}_Y\|^2)^{\frac{1}{2}}$. If an optimal coupling satisfies $\boldsymbol{Z}_X = \phi(\boldsymbol{Z}_Y)$ a.s. for some measurable function ϕ, we refer to ϕ as an optimal (transport) map from $\boldsymbol{X}$ to $\boldsymbol{Y}$. Optimal maps for the Wasserstein distance on the Euclidean line always exist and are explicitly available; on the contrary, on the Euclidean plane they are not guaranteed to exist if $\boldsymbol{X}$ gives non–zero mass to sets of codimension greater or equal to 1. Moreover, even when the existence is established, there is no explicit way to build such maps, except in few particular cases; see (Villani, 2008). However, Knott & Smith (1984) appointed a sufficient criterion to establish the optimality of a map, namely to express it as the gradient of a convex function. We will use this result in a reformulation provided by (Rüschendorf, 1991). When an optimal transport map ϕ is available, the Wasserstein distance amounts to an expected value with respect to a degenerate distribution having support on a 2–dimensional subspace of $\mathbb{R}^4$. Nonetheless, the evaluation of such an integral is still a difficult task since bivariate integrals can be difficult to evaluate not only analytically but also numerically.

3.3 Distance from exchangeability

Having established conditions for $d_{\mathcal{W}}$ in (3.1) to be a distance on completely random vectors, we now use $d_{\mathcal{W}}$ to compare CRVs $\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\xi}}$ in the same Fréchet class, i.e. with equally distributed marginal random measures ($\tilde{\mu}_1 \overset{\mathrm{d}}{=} \tilde{\xi}_1; \tilde{\mu}_2 \overset{\mathrm{d}}{=} \tilde{\xi}_2$), and focus on the comparison between their dependence structures. To this end, we put particular emphasis on the Wasserstein distance from comonotonic random vectors, which induce exchangeable priors. In this section, we provide an analytical expression for the optimal transportation map from a generic CRV to the comonotonic one in the same Fréchet class. This will then be used to evaluate the exact distance between exchangeability and the other extreme case, independence.

Definition 8. A random vector of measures $\tilde{\boldsymbol{\mu}}$ is said to be completely dependent or *comonotonic* if $\tilde{\mu}_1 = \tilde{\mu}_2$ a.s. We write $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}^{\mathrm{co}}$.

In particular, we point out that every random vector of measures $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \tilde{\mu}_2)$ in the same Fréchet class of $\tilde{\boldsymbol{\mu}}^{\mathrm{co}}$ satisfies $\tilde{\mu}_1 \overset{\mathrm{d}}{=} \tilde{\mu}_2$. For this reason, since our main interest lies in exchangeability and thus in comonotonicity, throughout the work we deal with random vectors of measures with equal marginal distributions. It should be stressed, though, that many of our results and techniques could be easily extended to other settings.

Theorem 20. *Let $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\mu}}^{\mathrm{co}}$ be CRVs in the same Fréchet class s.t. condition (3.6) on the Lévy intensities holds. Then,*

$$\mathcal{W}(\tilde{\boldsymbol{\mu}}(A), \tilde{\boldsymbol{\mu}}^{\mathrm{co}}(A))^2 = 4\left(\mathbb{E}\big(\tilde{\mu}_1(A)^2\big) - \omega_{\tilde{\boldsymbol{\mu}},A}\right), \tag{3.8}$$

where $\omega_{\tilde{\boldsymbol{\mu}},A} = \mathbb{E}\big(\tilde{\mu}_1(A)\, F^{-1}_{\tilde{\mu}_1(A)}(F_{\tilde{\mu}_1(A)+\tilde{\mu}_2(A)}(\tilde{\mu}_1(A)+\tilde{\mu}_2(A)))\big)$, with F_X denoting the cumulative distribution function (cdf) of X.

By defining $X_i = \tilde{\mu}_i(A)$ for $i = 1, 2$, it may be clarifying to observe that the right hand side of (3.8) is equal to $4(\mathbb{E}(X_1^2) - \mathbb{E}(X_1 F^{-1}_{X_1}(F_{X_1+X_2}(X_1+X_2))))$. In particular, when $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}^{\mathrm{co}}$, $X_1 = X_2 = X$, so that (3.8) becomes $4(\mathbb{E}(X^2) - \mathbb{E}(X F^{-1}_X(F_{2X}(2X)))) = 0$, since $F_{2X}(2X) = F_X(X)$. Moreover, when the distribution of $\tilde{\boldsymbol{\mu}}$ is symmetric, i.e. $\mathcal{L}(\tilde{\mu}_1, \tilde{\mu}_2) = \mathcal{L}(\tilde{\mu}_2, \tilde{\mu}_1)$, one finds the following alternative expression $\omega_{\tilde{\boldsymbol{\mu}},A}$ in (3.8).

Lemma 21. *Let $\tilde{\boldsymbol{\mu}}$ be a symmetric CRV satisfying the conditions of Theorem 20. Then*

$$\omega_{\tilde{\boldsymbol{\mu}},A} = \frac{1}{2}\mathbb{E}(F^{-1}_{\tilde{\mu}_1(A)+\tilde{\mu}_2(A)}(U) F^{-1}_{\tilde{\mu}_1(A)}(U)),$$

where $U \sim \mathrm{Unif}([0,1])$ is a uniform random variable on $[0,1]$.

Since $\tilde{\boldsymbol{\mu}}^{\mathrm{co}}$ is symmetric, we may apply Lemma 21 to check that $4\left(\mathbb{E}\big(\tilde{\mu}_1(A)^2\big) - \omega_{\tilde{\boldsymbol{\mu}},A}\right) = 0$ when $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}^{\mathrm{co}}$. Indeed this follows by observing that $F^{-1}_{2X}(U) F^{-1}_X(U) \overset{d}{=} 2X^2$, where $X = \tilde{\mu}_1(A) = \tilde{\mu}_2(A)$.
The expression of $\omega_{\tilde{\boldsymbol{\mu}},A}$ in Theorem 20 involves the dependence structure of $\tilde{\boldsymbol{\mu}}$ and is to

be evaluated case-by-case. In some specific cases it can be computed directly leading to the exact bivariate Wasserstein distance with respect to a comonotonic random vector in the same Fréchet class, in short *the Wasserstein distance from exchangeability*. For instance, consider a CRV $\tilde{\boldsymbol{\mu}}^{\text{ind}}$ whose marginals are independent gamma CRMs. Recall that $\tilde{\mu}$ is a gamma CRM with base measure $\alpha\, P_0$ if the Lévy intensity is

$$\pi(ds, dx) = \alpha\, P_0(dx)\, \frac{e^{-s}}{s} \mathbb{1}_{(0,+\infty)}(s)\, ds, \tag{3.9}$$

where $\alpha > 0$ and P_0 is a probability distribution on $\mathbb{X}$. This coincides with $\tilde{\mu} \sim \text{Ga}(1, \alpha P_0)$ defined in (1.6). We define

$$\omega_{\alpha,P_0,A} = \frac{1}{\Gamma(2\,\alpha\, P_0(A) + 1)} \int_0^{+\infty} \text{Inv}\Gamma_{\alpha\, P_0(A)}\left(\frac{\Gamma(\alpha\, P_0(A))}{\Gamma(2\,\alpha\, P_0(A))}\Gamma(2\,\alpha\, P_0(A), t)\right) e^{-t}\, t^{2\,\alpha\, P_0(A)}\, dt,$$

where $\Gamma(a, s) = \int_s^{+\infty} e^{-t}\, t^{a-1} dt$ is the upper incomplete gamma function and $\text{Inv}\Gamma_a(\cdot)$ is the inverse function of $\Gamma(a, \cdot)$.

Corollary 22. *Let $\tilde{\boldsymbol{\mu}}^{\text{ind}}$ and $\tilde{\boldsymbol{\mu}}^{\text{co}}$ be in the same Fréchet class with marginal gamma CRM with base measure $\alpha\, P_0$. Then,*

$$\mathcal{W}(\tilde{\boldsymbol{\mu}}^{\text{ind}}(A), \tilde{\boldsymbol{\mu}}^{\text{co}}(A))^2 = 4\,\alpha\, P_0(A)\,(1 + \alpha P_0(A) - \omega_{\alpha,P_0,A}).$$

Moreover,

$$\omega_{\alpha,P_0,A} = \frac{1}{2}\int_0^1 \text{Inv}\Gamma_{2\alpha\, \text{P}_0(\text{A})}(\text{t})\, \text{Inv}\Gamma_{\alpha\, \text{P}_0(\text{A})}(\text{t})\, \text{dt}.$$

For fixed values of $\alpha\, P_0(A)$, we can evaluate this quantity numerically. For example, Figure 3.1 corresponds to $\alpha = 1$ and $A = \mathbb{X}$, so that numerical simulations yield $\omega_{\alpha,P_0,A} \approx 1.19$. The analytical value is compared with the simulated Wasserstein distance between the empirical measures, which is known to converge to the Wasserstein distance between the underlying distributions as the size of the samples diverges. In many other cases the evaluation of the expression in Theorem 20 is impossible in practice. For example, this happens if the analytical expression for $F_{\tilde{\mu}_1(A)}$ is not available in closed form, or when the dependence between the random measures is modeled through the bivariate Lévy intensity. Moreover, we observe that the quantities in Theorem 20 and Corollary 22 depend on A in a non–trivial manner, so that finding the supremum over all Borel sets as in (3.1) may not be an easy task. This raises the need for informative and tractable upper bounds on the distance, whose expression depends directly on the underlying Lévy intensity. Note that the upper bound in (3.5) only depends on the marginal distributions of the random vectors, and thus does not provide any information on their dependence structures.

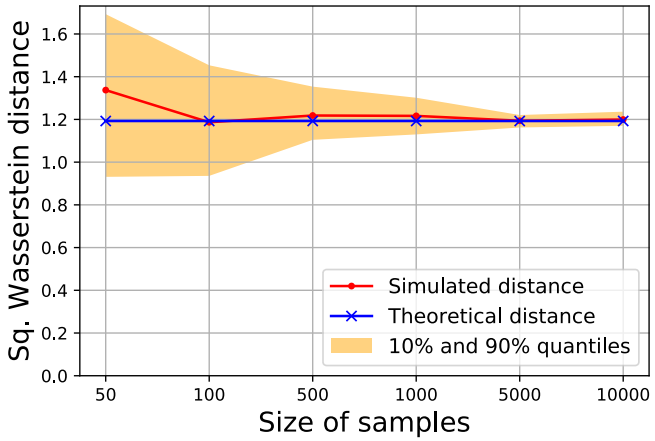

Figure 3.1: Simulation of the empirical Wasserstein distance between a bivariate distribution with independent gamma marginals with shape = scale = 1 and a bivariate distribution with a.s. equal gamma marginals of shape = scale = 1, based on 20 samples of increasing sizes. Simulations were performed with the Python Optimal Transport (POT) package (Flamary & Courty, 2017).

3.4 Bounds on Fréchet classes

Given the difficulty in evaluating the integral expression of Theorem 20 for the Wasserstein distance between a completely random vector and a comonotonic one in the same Fréchet class, we aim at deriving suitable bounds. We first face the problem in general and develop upper bounds for the Wasserstein distance between two CRVs. Then, in the following sections, these general bounds will be specialized to the distance from exchangeability, which is the case of interest for Bayesian inference. Our general bounds rely on a compound Poisson approximation of the CRVs, which are induced by certain *compatible* families of neighborhoods of the origin. Henceforth we assume that $\tilde{\boldsymbol{\mu}}$ is an infinitely active CRV s.t. condition (3.6) on the Lévy intensity ν holds.

Definition 9. Consider a family $B = \{B(\epsilon) \,|\, \epsilon \in (0,1]\}$ of measurable neighborhoods of the origin in $\mathbb{R}^2_+$ s.t.

(B1) the family is increasing, i.e. $\epsilon_1 \leq \epsilon_2$ implies that $B(\epsilon_1) \subset B(\epsilon_2)$;

(B2) the Lévy intensity gives zero mass to their intersection, i.e. $\nu(\cap_{\epsilon\in(0,1]} B(\epsilon) \times A) = 0$ for every $A \in \mathcal{X}$;

(B3) the sets $D = \{D(\epsilon) = B(\epsilon)^c = \mathbb{R}^2_+ \setminus B(\epsilon) \,|\, \epsilon \in (0,1]\}$ have continuously increasing mass, i.e. there exists $r_0 = \nu(\cap_{\epsilon\in(0,1]} D(\epsilon))$ s.t. for every $r > r_0$ there exists $\epsilon_r = \epsilon_{r,A}$ s.t. $\nu(D(\epsilon_r) \times A) = r$.

Then we say that the family B is *compatible* with $\tilde{\boldsymbol{\mu}}$. By extension, we will also refer to the family of complementary sets D as compatible.

Remark 2. Some technical comments are in order: (a) The choice of the index set to be $(0, 1]$ is arbitrary. Indeed, one could replace it with any neighborhood of the origin in $\mathbb{R}^+$. (b) The uncountable intersection $\cap_{\epsilon\in(0,1]} D(\epsilon)$ is measurable because the family is increasing. One can find more on this in Section 3.9. (c) Property (B2) does not contradict the continuity of the measure since ν is an infinite measure.

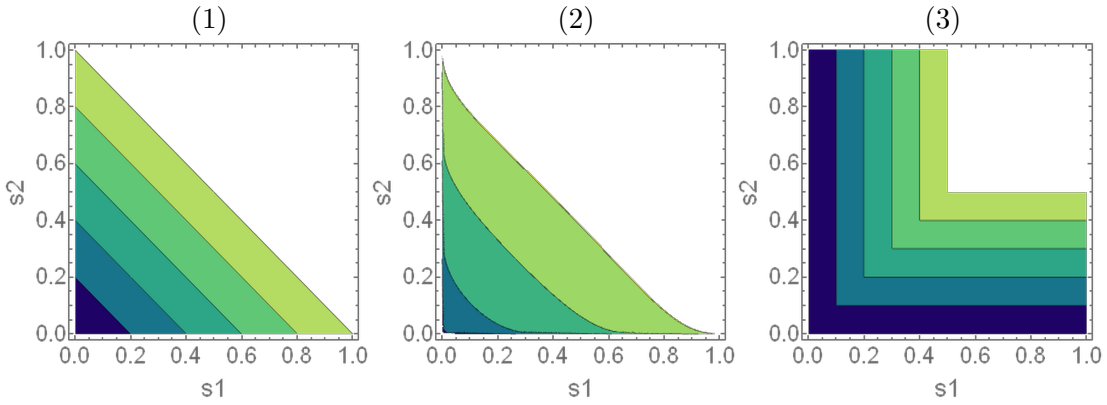

Figure 3.2: Three families of neighborhoods of the origin:
(1) $B^+(\epsilon) = \{(s_1, s_2) \,|\, s_1 + s_2 \leq \epsilon\}$;
(2) $B^{E_1}(\epsilon) = \{(s_1, s_2) \,|\, E_1(s_1)^{-\theta} + E_1(s_2)^{-\theta} \leq \epsilon\}$ with $\theta = 0.5$;
(3) $B^{\min}(\epsilon) = \{(s_1, s_2) \,|\, \min(s_1, s_2) \leq \epsilon\}$.

Remark 3. A standard way to find a family of measurable neighborhoods of the origin that satisfy (B1) is to consider the level sets

$$B^g(\epsilon) = \{(s_1, s_2) \,|\, g(s_1, s_2) \leq \epsilon\}, \tag{3.10}$$

where $g : [0, +\infty) \times [0, +\infty) \to \mathbb{R}^+$ is a measurable function s.t. $g(0, 0) = 0$. See Figure 3.2. Depending on the support of ν, properties (B2) and (B3) may hold. For example, if $g(s_1, s_2) = \min(s_1, s_2)$, B^g is not compatible with ν having mass on the axis, whereas it is compatible with ν being absolutely continuous (a.c.) w.r.t. the Lebesgue measure.
As will be seen in the sequel, we will mostly be interested in Lévy intensities that are a.c. w.r.t. the Lebesgue measure or have mass on lines passing through the origin. In these cases, every continuous map g s.t. $g(s_1, s_2) = 0$ if and only if $(s_1, s_2) = (0, 0)$ induces a compatible family. In particular, we will be interested in the families B^+ and B^{E_1} appearing in Figure 3.2, where $E_1(s) = \Gamma(0, s)$ is the exponential integral.

Given a compatible family D, for every $r > r_0$ and every $A \in \mathcal{X}$ we define the probability distribution $\rho_{r,A,D}$ on $\mathbb{R}^2_+$ as

$$\rho_{r,A,D}(ds_1, ds_2) = \frac{1}{r}\, \nu(ds_1, ds_2, A)\, \mathbb{1}_{D(\epsilon_{r,A})}(s_1, s_2), \tag{3.11}$$

where we have used the notation $\nu(ds_1, ds_2, A) = \int_A \nu(ds_1, ds_2, dy)$. As apparent from the proof of the next theorem, this coincides with the distribution of the jumps of a compound Poisson approximation of $\tilde{\boldsymbol{\mu}}$.

Theorem 23. *Let $\tilde{\boldsymbol{\mu}}^1$ and $\tilde{\boldsymbol{\mu}}^2$ be infinitely active CRVs in the same Fréchet class s.t. condition (3.6) on the Lévy intensities holds. Then,*

$$\mathcal{W}(\tilde{\boldsymbol{\mu}}^1(A), \tilde{\boldsymbol{\mu}}^2(A)) \leq \lim_{r \to +\infty} \sqrt{r}\ \mathcal{W}(\rho^1_{r,A,D_1}, \rho^2_{r,A,D_2}),$$

for any D_i compatible family for $\tilde{\boldsymbol{\mu}}^i$, for $i = 1, 2$. Moreover, the upper bound on the right hand side is finite and does not depend on D_1 and D_2.

Remark 4. Since any completely random vector $\tilde{\boldsymbol{\mu}}$ has infinitely many compatible family, the above theorem holds also in the case $\tilde{\boldsymbol{\mu}}^1 = \tilde{\boldsymbol{\mu}}^2$ and $D_1 \neq D_2$. Since the limit does not depend on the families D_1 and D_2, we know that in such case it is equal to zero.

The proof is detailed in Section 3.9 and is based on a bound on the Wasserstein distance between compound Poisson distributions. A similar problem was treated in (Mariucci & Reiß, 2018) for Lévy processes on $\mathbb{R}$. Nonetheless, the extension to $\mathbb{R}^2$ needs a new bound on the compound Poisson distributions in $\mathbb{R}^2$, summarized by the following proposition. Indeed, the arguments used in (Mariucci & Reiß, 2018, Theorem 10) could be used to bound the Wasserstein distance from above with $\sqrt{r + r^2}\ \mathcal{W}(\rho^1_{r,A,D_1}, \rho^2_{r,A,D_2})$, which goes to $+\infty$ as $r \to +\infty$.

Proposition 24. *Let $\boldsymbol{X} \overset{d}{=} \sum_{i=1}^{N_x} \boldsymbol{X}^i$ and $\boldsymbol{Y} \overset{d}{=} \sum_{i=1}^{N_y} \boldsymbol{Y}^i$ be two compound Poisson processes in $\mathbb{R}^2$ s.t. N_x and N_y are Poisson random variables with mean r and $\{\boldsymbol{X}^i \,|\, i \geq 1\}$ and $\{\boldsymbol{Y}^i \,|\, i \geq 1\}$ are sequences of independent and identically distributed random elements in $\mathbb{R}^2$, independent from N_x and N_y respectively. Then*

$$\mathcal{W}(\boldsymbol{X}, \boldsymbol{Y})^2 \leq r\mathcal{W}(\boldsymbol{X}^1, \boldsymbol{Y}^1)^2 + (r^2 - r)\, \|\mathbb{E}(\boldsymbol{X}^1) - \mathbb{E}(\boldsymbol{Y}^1)\|^2.$$

Remark 5. Theorem 23 bounds the Wasserstein distance between the completely random vectors with the Wasserstein distance between quantities that only depend on the bivariate Lévy intensities. Yet, the Wasserstein distance between these two quantities suffers from all the technical difficulties related to the Wasserstein distance in $\mathbb{R}^2$. Hence, it is complicated to evaluate it, analytically and numerically. The next sections are devoted to this task.

3.5 Bounds on exchangeability

Our next goal is to measure the dependence of a given completely random vector as the Wasserstein distance from exchangeability, which is induced by comonotonic CRVs. For this reason, we now specialize the results of the previous section, which apply to all completely random vectors in the same Fréchet class, to this particular framework of

great importance for Bayesian inference.
In order to evaluate the bound in Theorem 23 numerically, we first need an explicit expression for the Wasserstein distance between the jumps of the compound Poisson approximations. With this goal in mind, we first dwell on the Lévy intensity ν^{co} of a comonotonic random vector $\tilde{\boldsymbol{\mu}}^{\text{co}}$.

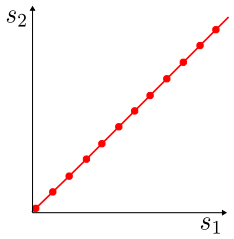

Figure 3.3: Support of the Lévy intensity of a comonotonic completely random vector.

Proposition 25. *For every $A \in \mathcal{X}$, the Lévy intensity $\nu^{\text{co}}(ds_1, ds_2, A)$ has support on the bisecting line of $\mathbb{R}^2_+$, i.e.*

$$\nu^{\text{co}}(ds_1, ds_2, A) = \delta_{s_1}(ds_2)\, \nu_1(ds_1, A) = \delta_{s_2}(ds_1)\, \nu_2(ds_2, A).$$

It follows that every random vector $\tilde{\boldsymbol{\mu}}$ in the same Fréchet class of $\tilde{\boldsymbol{\mu}}^{\text{co}}$ has equal marginal Lévy intensities $\nu_1(ds, dx) = \nu_2(ds, dx)$, which we denote with $\pi(ds, dx)$. In particular for every $A \in \mathcal{X}$, $\pi(ds, A) = \pi(s, A)\, ds$ is a.c. w.r.t. the Lebesgue measure and infinitely active. We denote with $U^{\pi}_A(t) = \int_{[t,+\infty)} \pi(s, A)\, ds$ its tail integral.
The following theorem provides the exact expression of the limit appearing in Theorem 23 together with a class of upper bounds. The latter are useful when the exact expression cannot be evaluated analytically or numerically, as will be seen in Section 3.7.2. We first define some relevant quantities:

$$h^g_{\nu,A}(s) = \int_{\mathbb{R}^2_+} \mathbb{1}_{(s,+\infty)}(g(t_1, t_2))\, \nu(dt_1, dt_2, A); \tag{3.12}$$

$$K^g_{\nu,A} = \sum_{i=1}^{2} \int_{\mathbb{R}^2_+} |s_i - (U^{\pi}_A)^{-1}(h^g_{\nu,A}(g(s_1, s_2)))|^2\, \nu(ds_1\, ds_2, A);$$

where $g : \mathbb{R}^2 \to \mathbb{R}$ is a measurable map. When $g(s_1, s_2) = s_1 + s_2$ we write $h^+_{\nu,A}$ and $K^+_{\nu,A}$. In particular, since $g(s_1, s_2) = s_1 + s_2$ is symmetric and ν has equal marginal measures

$$K^+_{\nu,A} = 2 \int_{\mathbb{R}^2_+} |s_i - (U^{\pi}_A)^{-1}(h^+_{\nu,A}(s_1 + s_2))|^2\, \nu(ds_1\, ds_2, A). \tag{3.13}$$

Theorem 26. *Let $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\mu}}^{\mathrm{co}}$ satisfy the conditions of Theorem 23 s.t. B^+ defined in Remark 3 is compatible with $\tilde{\boldsymbol{\mu}}$. Then*

$$\lim_{r\to+\infty} r\,\mathcal{W}(\rho_{r,A,D}, \rho^{\mathrm{co}}_{r,A,D^{\mathrm{co}}})^2 = K^+_{\nu,A}. \tag{3.14}$$

Moreover, for every continuously differentiable $g:\mathbb{R}^2\to\mathbb{R}$ s.t. B^g is compatible with $\tilde{\boldsymbol{\mu}}$, $K^+_{\nu,A} \le K^g_{\nu,A}$.

Theorem 26 thus establishes that $g(s_1,s_2)=s_1+s_2$ realizes the optimal bound in the class $\{K^g_{\nu,A}\}$. The expression for $K^+_{\nu,A}$ resembles the one for the Wasserstein distance in Theorem 20 and is derived in a similar way. Nonetheless, by working at the level of the bivariate Lévy intensities rather than at the level of the one–dimensional distributions $\tilde{\boldsymbol{\mu}}(A)$, we overcome many of the difficulties related to its evaluation. In particular, when the Lévy intensity $\nu(\,\cdot\,,A)$ is a.c. w.r.t. the Lebesgue measure on $\mathbb{R}^2$ for any A in $\mathcal{X}$, $K^+_{\nu,A}$ comes in a compelling form. In a such case we denote with $\nu(s_1,s_2,A)$ its Radon–Nikodym derivative and define

$$K_{\nu,A} = \int_0^{+\infty} (U^\pi_A)^{-1}(h^+_{\nu,A}(t)) \int_0^t s\,\nu(s,t-s,A)\,ds\,dt,$$

where $h^+_{\nu,A}$ is as in (3.12).

Theorem 27. *Let $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\mu}}^{\mathrm{co}}$ satisfy the conditions of Theorem 23. If the Lévy intensity of $\tilde{\boldsymbol{\mu}}$ is such that, for any $A\in\mathcal{X}$, $\nu(\,\cdot\,,A)$ is a.c. w.r.t. the Lebesgue measure on $\mathbb{R}^2$, then*

$$\lim_{r\to+\infty} r\,\mathcal{W}(\rho_{r,A,D}, \rho^{\mathrm{co}}_{r,A,D^{\mathrm{co}}})^2 = 4\left(\int_0^{+\infty} s^2\,\pi(ds,A)\,ds - K_{\nu,A}\right)$$

Remark 6. We observe that the first integral in the bound only depends on the marginal distributions and provides a general upper bound for the distance. This can be seen as an improvement of the bound in (3.5), which amounts to

$$\mathcal{W}(\tilde{\boldsymbol{\mu}}(A), \tilde{\boldsymbol{\mu}}^{\mathrm{co}}(A))^2 \le 4\left(\int_0^{+\infty} s^2\,\pi(s,A)\,ds + \int_0^{+\infty} s\,\pi(s,A)\,ds\right),$$

where π is the marginal Lévy intensity, as defined at the beginning of the section. On the other hand, $K_{\nu,A}$ provides information contained in the dependence structure. In Section 3.7.1 this will be specialized for a concrete example.

Remark 7. When the Lévy intensities are *homogeneous*, i.e.

$$\nu(ds_1,ds_2,dx) = \alpha\,P_0(dx)\,\nu(ds_1,ds_2) \tag{3.15}$$

where P_0 is a probability distribution on $\mathbb{X}$ and $\alpha>0$, also the marginal Lévy intensity takes the form $\pi(dx,ds)=\alpha\,P_0(dx)\,\pi(s)\,ds$ and we denote by $U_\pi(t)=\int_t^{+\infty}\pi(s)\,ds$ the tail integral. If the Lévy intensity is also diffuse, $K^+_{\nu,A}=\alpha\,P_0(A)\,K_\nu$, where

$$\begin{aligned} K_\nu &= \int_0^{+\infty} (U^\pi)^{-1}(h^+_\nu(t)) \int_0^t s\,\nu(s,t-s)\,ds\,dt; \\ h^+_\nu(s) &= \int_{\mathbb{R}^2_+} \mathbb{1}_{(s,+\infty)}(t_1+t_2)\,\nu(t_1,t_2)\,dt_1\,dt_2. \end{aligned} \tag{3.16}$$

In particular, this entails that

$$d_{\mathcal{W}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}^{\mathrm{co}})^2 \leq 4\,\alpha \left(\int_0^{+\infty} s^2\, \pi(ds)\, ds - K_\nu \right).$$

3.6 Independence

In this section we will use Proposition 24 to bound the distance between exchangeability and the other extreme case, independence. As we shall see, in this case the Lévy intensity is not a.c. w.r.t. the Lebesgue measure and thus the results of Theorem 27 do not apply.

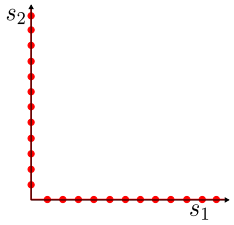

Figure 3.4: Support of the Lévy intensities of a completely random vector with independent marginals.

Let $\tilde{\boldsymbol{\mu}}^{\mathrm{ind}}$ be completely random vector with independent marginals and let ν^{ind} denote its Lévy intensity. An immediate adaptation of (Kallsen & Tankov, 2006, Lemma 4.1) shows that the corresponding Lévy intensities $\nu^{\mathrm{ind}}(ds_1, ds_2, A)$ have support on the axis, namely

$$\nu^{\mathrm{ind}}(ds_1, ds_2, A) = \delta_0(ds_2)\ \nu_1^{\mathrm{ind}}(ds_1, A) + \delta_0(ds_1)\ \nu_2^{\mathrm{ind}}(ds_2, A).$$

In our setting, $\nu_1^{\mathrm{ind}}(ds_1, A) = \nu_2^{\mathrm{ind}}(ds_2, A) = \pi(s, A)\, ds$. Before stating the main result, we introduce the following quantity, which only depends on the marginal distribution π of the completely random vectors:

$$K_{\pi,A} = \int_0^{+\infty} (U_A^\pi)^{-1}(2\, U_A^\pi(s))\, s\, \pi(s, A)\, ds.$$

Theorem 28. *Let $\tilde{\boldsymbol{\mu}}^{\mathrm{ind}}$ and $\tilde{\boldsymbol{\mu}}^{\mathrm{co}}$ be in the same Fréchet class s.t. the conditions of Theorem 23 hold. Then*

$$\lim_{r \to +\infty} r\, \mathcal{W}(\rho_{r,A}^{\mathrm{ind}}, \rho_{r,A}^{\mathrm{co}})^2 = 4 \left(\int_0^{+\infty} s^2\, \pi(s, A)\, ds - K_{\pi,A} \right).$$

Remark 8. Similarly to Remark 7, when the Lévy intensities are homogeneous, $K_{\pi,A} = \alpha\, P_0(A)\, K_\pi$, where $K_\pi = \int_0^{+\infty} (U^\pi)^{-1}(2\, U^\pi(s))\, s\, \pi(s)\, ds$.

We now apply Theorem 28 to the case where the marginal distribution is a gamma CRM with base measure $\alpha\, P_0$, as in Corollary 22, which allows us to compare the exact

Wasserstein distance with the relative bound. We first define the constant

$$\gamma = 4 - 4 \int_0^{+\infty} (E_1)^{-1}(2\, E_1(s))\, e^{-s}\, ds, \tag{3.17}$$

where, as before, $E_1(s) = \Gamma(0, s)$ is the exponential integral. Numerical integration show that $\gamma \approx 1.24$.

Corollary 29. *Let $\tilde{\boldsymbol{\mu}}^{\text{ind}}$ and $\tilde{\boldsymbol{\mu}}^{\text{co}}$ be in the same Fréchet class with marginal gamma CRM with base measure $\alpha\, P_0$. Then,*

$$\mathcal{W}(\tilde{\boldsymbol{\mu}}^{\text{ind}}(A), \tilde{\boldsymbol{\mu}}^{\text{co}}(A))^2 \leq \gamma\, \alpha\, P_0(A).$$

In particular, $d_{\mathcal{W}}(\tilde{\boldsymbol{\mu}}^{\text{ind}}, \tilde{\boldsymbol{\mu}}^{\text{co}})^2 \leq \gamma\, \alpha$.

In Figure 3.5 we present a graphical comparison between the exact distance in Corollary 22, the simulated empirical distance in Figure 3.1 as the sample size increases and the theoretical bound established in Theorem 29. We omit the non–informative bound in Remark 6 from the figure because it is out of scale (equal to 8) and point out that the theoretical bound appears to be very tight.

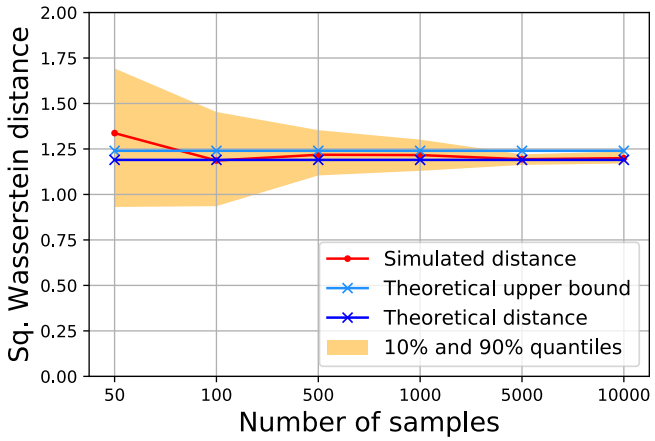

Figure 3.5: Simulations of the Wasserstein distance in Figure 3.1 compared with the non–informative bound in Remark 6 and the informative bound in Theorem 29. Simulations were performed with a growing sample size using the Python Optimal Transport (POT) package (Flamary & Courty, 2017).

Similar results may be achieved for *generalized gamma* completely random measures, as defined in (1.8) in Section 1.4, whose Lévy measure is

$$\pi(ds, dx) = \alpha\, P_0(dx)\, e^{-b\, s}\, s^{-1-\sigma} \mathbb{1}_{(0,+\infty)}(s)\, ds,$$

for some $\alpha > 0$, P_0 a probability distribution on $\mathbb{X}$, $b > 0$ and $\sigma \in (0, 1)$. In particular, gamma random measures as defined in (3.9) are achieved when $\sigma = 0$ and $b = 1$. We

define

$$\gamma_{b,\sigma} = 4 - 4\frac{1}{b\,\Gamma(1-\sigma)}\int_0^{+\infty} \mathrm{Inv}\Gamma_{-\sigma}(2\,\Gamma(-\sigma, b\,s))\, e^{-b\,s}\, s^{-\sigma}\, ds, \tag{3.18}$$

where $\Gamma(a,s) = \int_s^{+\infty} e^{-t}\, t^{a-1} dt$ is the upper incomplete gamma function and $\mathrm{Inv}\Gamma_a(\cdot)$ is the inverse function of $\Gamma(a,\cdot)$. Clearly, $\gamma_{1,0} = \gamma$ in (3.17).

Corollary 30. *Let $\tilde{\boldsymbol{\mu}}^{\mathrm{ind}}$ and $\tilde{\boldsymbol{\mu}}^{\mathrm{co}}$ be in the same Fréchet class with marginal generalized gamma CRM with parameters b, σ and base measure $\alpha\, P_0$. Then,*

$$\mathcal{W}(\tilde{\boldsymbol{\mu}}^{\mathrm{ind}}(A), \tilde{\boldsymbol{\mu}}^{\mathrm{co}}(A))^2 \leq \gamma_{b,\sigma}\, \alpha\, P_0(A).$$

In particular, $d_{\mathcal{W}}(\tilde{\boldsymbol{\mu}}^{\mathrm{ind}}, \tilde{\boldsymbol{\mu}}^{\mathrm{co}})^2 \leq \gamma_{b,\sigma}\, \alpha$.

The bounds in Corollary 30 shed light on the role of the hyperparameters in the distance from exchangeability. In particular, Figure 3.6 shows that the distance increases linearly as σ increases and logarithmically as b increases.

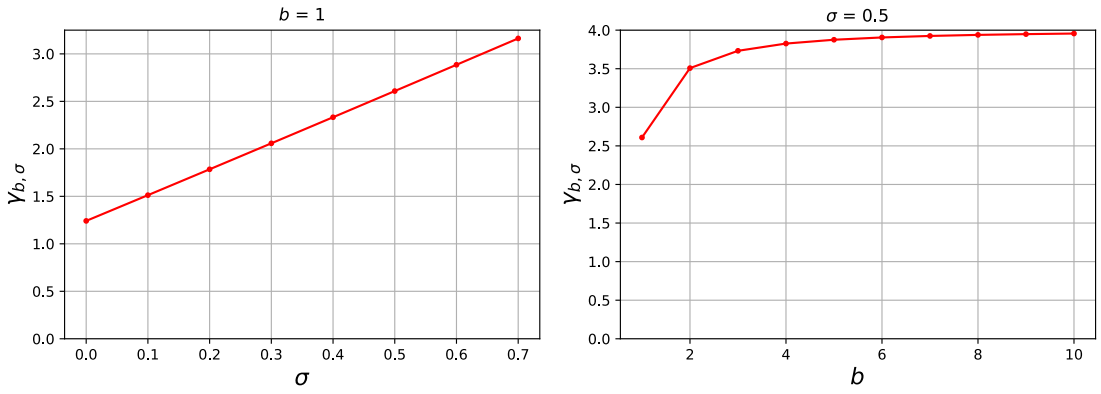

Figure 3.6: Numerical integrations of $\gamma_{b,\sigma}$. On the left, $b = 1$ and σ varies from 0 to 0.7. On the right, $\sigma = 0.5$ while b varies from 1 to 10.

3.7 Measuring dependence in BNP models

We now analyze three popular procedures to model the dependence between CRMs through the choice of an hyperparameter, namely compound random measures, Clayton–Lévy copula and GM–dependence. These can be seen as the infinite–dimensional extension of the *approximately* exchangeable priors suggested by de Finetti (de Finetti, 1938) for binary data, and further investigated in Bacallado et al. (2015). Our theoretical findings allow for a formal quantification of the dependence in terms of a meaningful bound on the distance from exchangeability. These bounds are expressed in terms of the models' hyperparameters leading to intuitive results, which can also guide the parameters' elicitation.

3.7.1 Compound random measures

Compound random measures, introduced in Griffin & Leisen (2017), provide a general framework for building completely random vectors. As underlined in Section 1.6.4, these may be used to model the dependence between CRMs with many different marginal distributions, such as gamma, generalized gamma, beta and σ-stable random measures.

Definition 10. A *compound random measure* $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \tilde{\mu}_2)$ is a completely random vector of the form

$$\begin{pmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_2 \end{pmatrix} = \sum_{i=1}^{+\infty} \begin{pmatrix} m_{1,i} \\ m_{2,i} \end{pmatrix} J_i \, \delta_{X_i},$$

where $\tilde{\eta} = \sum_{i=1}^{+\infty} J_i \, \delta_{X_i}$ is a homogeneous CRM with Lévy intensity $\alpha \, P_0(dx) \, \nu^*(ds)$ and $(m_{1,i}, m_{2,i}) \overset{\text{iid}}{\sim} h$, where h is a bivariate density.

In Griffin & Leisen (2017) the authors prove that such $\tilde{\boldsymbol{\mu}}$ is a completely random vector with bivariate Lévy intensity

$$\nu(ds_1, ds_2, dx) = \alpha \, P_0(dx) \int_{\mathbb{R}^+} \frac{1}{u^2} \, h\left(\frac{s_1}{u}, \frac{s_2}{u}\right) \nu^*(du) \, ds_1, ds_2.$$

Specific choices for ν^* and h lead to different marginal CRMs and dependence structures. In particular, by taking h corresponding to the distribution of two independent gamma $(\phi, 1)$ random variables and $\nu^*(du) = (1-u)^{\phi-1} u^{-1} \mathbb{1}_{(0,1)}(u) \, du$, one achieves marginal gamma random measures of shape parameter 1 and base measure $\alpha \, P_0$. We write $\tilde{\boldsymbol{\mu}} \sim \text{CoGamma}(\phi, \alpha, P_0)$. Here we focus on the case of gamma marginal random measures, though the techniques may be generalized. Our aim is to quantify dependence, which is controlled by the parameter ϕ. We first introduce some relevant quantities.

$$K_\phi = \int_0^{+\infty} E_1^{-1}(e(\phi, t)) \, \phi \, f(\phi, 2\,\phi, t) \, dt,$$

$$e(\phi, t) = \frac{1}{\Gamma(2\,\phi)} \int_0^1 \Gamma\left(2\,\phi, \frac{t}{u}\right) (1-u)^{\phi-1} \, u^{-1} \, du, \quad e_{\mathbb{N}}(\phi, t) = \sum_{k=0}^{2\phi-1} f(\phi, k, t)$$

$$f(\phi, x, t) = \frac{t^x}{\Gamma(x)} \int_0^1 e^{-\frac{t}{u}} (1-u)^{\phi-1} \, u^{-x-1} \, du.$$

$$f_{\mathbb{N}}(\phi, n, t) = \frac{t^n}{n!} \sum_{j=0}^{\phi-1} \binom{\phi-1}{j} (-1)^j g(n, j, t),$$

where $g(n, j, t)$ is equal to

$$\begin{cases} t^{-n+j} \, (n-j-1)! \, e^{-t} \, \sum_{h=0}^{n-j-1} \frac{t^h}{h!} & \text{if } n > j \\ \frac{1}{(j-n)!} \left(e^{-t} \, \sum_{j=0}^{j-n-1} (-1)^h \, (j-n-h-1)! t^h + (-1)^{j-n} \, E_1(t)\right) & \text{if } n \leq j. \end{cases}$$

Theorem 31. *Let* $\tilde{\boldsymbol{\mu}} \sim \mathrm{CoGamma}(\phi, \alpha, \mathrm{P}_0)$ *and let* $\tilde{\boldsymbol{\mu}}^{\mathrm{co}}$ *denote the comonotonic random vector in the same Fréchet class. Then,*

$$\mathcal{W}(\tilde{\boldsymbol{\mu}}(A), \tilde{\boldsymbol{\mu}}^{\mathrm{co}}(A))^2 \leq 4\,\alpha\,P_0(A)\,(1 - K_\phi).$$

In particular, $d_{\mathcal{W}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}^{\mathrm{co}})^2 \leq 4\,\alpha\,(1 - K_\phi)$. *Moreover, when* $\phi \in \mathbb{N}$, $e = e_{\mathbb{N}}$ *and* $f = f_{\mathbb{N}}$.

Theorem 31 allows to conveniently compute the Wasserstein distance from exchangeability for ϕ an integer value. Table 3.1 displays some numerical results for different values of ϕ. As ϕ increases, the dependence between the induced marginal gamma random measures also increases. Moreover, we stress that the case $\phi = 1$ is of particular interest since it corresponds to the dependence structure discussed in (Leisen et al., 2013).

ϕ	$1 - K_\phi$ $(\approx)$
1	0.1426
5	0.0545
10	0.0241
30	0.0081

Table 3.1: Values of the constant $1 - K_\phi$ appearing in the bound of Theorem 31 for different values of ϕ.

We may compare the theoretical upper bounds in Theorem 31 with the simulated Wasserstein distance, as in Figure 3.4 and 3.8. As in the previous cases, our upper bounds appear to be tight and informative.

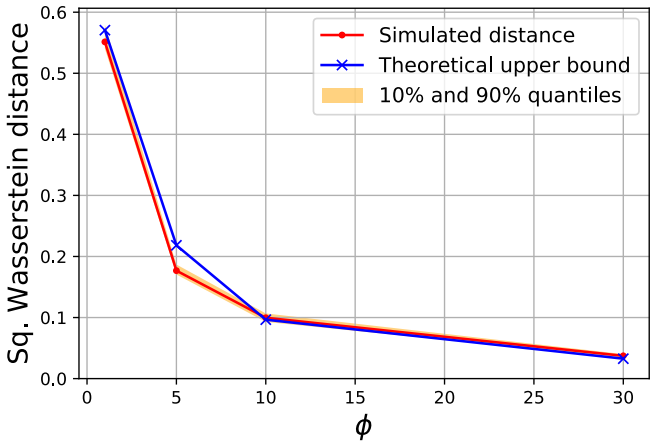

Figure 3.7: Simulation of the Wasserstein distance between a random vector $(\mu_1(\mathbb{X}), \mu_2(\mathbb{X}))$ with marginal compound random measures of parameters (ϕ, α, P_0), where $\alpha = 1$ and ϕ varies, and a bivariate distribution with a.s. equal gamma marginals of shape = scale = 1. The simulations were performed with samples of 10000 observations.

3.7.2 Clayton–Lévy copula

This last section focuses on Lévy copulae, which provide another popular way to model dependence between completely random measures, by separating the marginal components of a bivariate Lévy measure from its dependence structure. Lévy copulae were introduced in Tankov (2003) and Cont & Tankov (2004) to model the dependence structure between Lévy processes, and have been further used on completely random measures (Epifani & Lijoi, 2010; Leisen & Lijoi, 2011). For details we refer to Section 1.6.5 and Kallsen & Tankov (2006).

Lévy copulae provide another popular way to model dependence between CRMs. Standard copulae can be seen as a means to separate the marginal components of a bivariate distribution from its dependence structure. The same happens for their generalization to Lévy intensities, conceived in Tankov (2003) and Cont & Tankov (2004) to model the dependence structure between Lévy processes. See also Section 1.6.5 and (Kallsen & Tankov, 2006; Epifani & Lijoi, 2010; Leisen & Lijoi, 2011) for uses on CRMs. Given a bivariate Lévy intensity $\nu(ds_1, ds_2, A)$, we indicate by $U_{i,A}(t) = \int_t^\infty \nu_i(ds, A)$, for $i = 1, 2$, its marginal tail integrals. An analogue of Sklar's Theorem states that there exists a Lévy copula $c : [0, +\infty]^2 \to [0, +\infty]$ s.t.

$$\nu((t_1, +\infty) \times (t_2, +\infty) \times A) = c(U_{1,A}(t_1), U_{2,A}(t_2)).$$

When the Lévy copula c and the tail integrals $U_{1,A}, U_{2,A}$ are sufficiently smooth, $\nu(ds_1, ds_2, A)$ is recovered by

$$\nu(ds_1, ds_2, A) = \frac{\partial^2}{\partial u_1 \partial u_2} c(u_1, u_2)\big|_{U_{1,A}(s_1), U_{2,A}(s_2)} \nu_1(ds_1, A)\, \nu_2(ds_2, A). \tag{3.19}$$

It follows that Lévy copulae are useful to build bivariate Lévy intensities, allowing to gain insight into their dependence structure. Consider the Clayton–Lévy copula, which is a smooth class of copulae with both independence and complete dependence as limiting cases:

$$c_\theta(s_1, s_2) = (s_1^{-\theta} + s_2^{-\theta})^{-\frac{1}{\theta}},$$

for $\theta > 0$. This was used, for example, in (Epifani & Lijoi, 2010; Leisen & Lijoi, 2011). As $\theta \to +\infty$ one achieves the complete dependence copula (Kallsen & Tankov, 2006) which, by taking equal marginal Lévy intensities, corresponds to the exchangeability assumption. We write $\tilde{\boldsymbol{\mu}} \sim \mathrm{Cl}(\theta, \alpha, P_0)$ for a completely random vector with marginal gamma random measures with base measure αP_0 and Lévy copula c_θ. Our goal is to show that, as $\theta \to +\infty$, $\tilde{\boldsymbol{\mu}}$ converges in the Wasserstein distance to the comonotonic random vector with same marginal distributions and also to provide an upper bound for the rate of convergence. Define

$$K_\theta = \frac{1+\theta}{\theta^2} \int_0^\infty \int_{y_1}^\infty E_1^{-1}\big(y_1^{-\frac{1}{\theta}}\big)\, E_1^{-1}\left(\frac{1+\theta}{\theta} y_2^{-\frac{1}{\theta}}\right) y_2^{-\frac{1}{\theta}-2} dy_1\, dy_2.$$

Theorem 32. *Let $\tilde{\boldsymbol{\mu}} \sim \mathrm{Cl}(\theta, \alpha, P_0)$ and let $\tilde{\boldsymbol{\mu}}^{\mathrm{co}}$ be in the same Fréchet class. Then*

$$d_{\mathcal{W}}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}^{\mathrm{co}})^2 \leq 4\,\alpha\,(1 - K_\theta).$$

Moreover, as $\theta \to +\infty$, K_θ goes to 1.

3.7.3 GM–dependence

In the next nonparametric model we consider, introduced in (Lijoi et al., 2014; Lijoi & Nipoti, 2014), the dependence between CRMs is induced by the bivariate Poisson process proposed in Griffiths & Milne (1978), which brings to the appealing additive structure anticipated in Section 1.6.3.

Definition 11. A completely random vector $\boldsymbol{\xi}$ is GM–dependent if

$$\begin{pmatrix} \tilde{\xi}_1 \\ \tilde{\xi}_2 \end{pmatrix} \stackrel{\mathrm{d}}{=} \begin{pmatrix} \tilde{\mu}_1 + \tilde{\mu}_0 \\ \tilde{\mu}_2 + \tilde{\mu}_0 \end{pmatrix}, \tag{3.20}$$

where $\tilde{\mu}_0, \tilde{\mu}_1$ and $\tilde{\mu}_2$ are three independent CRMs with Lévy intensities

$$\begin{aligned} v_1(ds, dx) = v_2(ds, dx) &= \alpha\, z\, P_0(dx)\, \rho(s)\, ds \\ v_0(ds, dx) &= \alpha\, (1 - z)\, P_0(dx)\, \rho(s)\, ds, \end{aligned}$$

where $\alpha > 0$, $z \in (0, 1)$, P_0 is a probability measure on $\mathbb{R}$ and ρ is a measurable function.

Set $\tilde{\boldsymbol{\mu}}^{\mathrm{ind}} = (\tilde{\mu}_1, \tilde{\mu}_2)$ and $\tilde{\boldsymbol{\mu}}_0^{\mathrm{co}} = (\tilde{\mu}_0, \tilde{\mu}_0)$ to underline that they are, respectively, an independent and a comonotonic completely random vector. The completely random vector $\boldsymbol{\xi}$ has marginal Lévy intensity $\pi(ds, dx) = \alpha\, P_0(dx)\, \rho(s)\, ds$, but we are not given the corresponding bivariate Lévy intensity. Nonetheless, the next result provides bounds on its distance from the comonotonic and the random vector with independent marginals in the same Fréchet class, in terms of the underlying random vectors $\tilde{\boldsymbol{\mu}}^{\mathrm{ind}}, \tilde{\boldsymbol{\mu}}_0^{\mathrm{co}}$.

Proposition 33. *Let $\tilde{\boldsymbol{\xi}}$ be a* GM*–dependent CRV and let $\tilde{\boldsymbol{\xi}}^{\mathrm{co}}$ denote the comonotonic random vector in the same Fréchet class. Then*

$$\begin{aligned} d_{\mathcal{W}}(\tilde{\boldsymbol{\xi}}, \tilde{\boldsymbol{\xi}}^{\mathrm{co}}) &\leq d_{\mathcal{W}}(\tilde{\boldsymbol{\mu}}^{\mathrm{ind}}, \tilde{\boldsymbol{\mu}}^{\mathrm{co}}); \\ d_{\mathcal{W}}(\tilde{\boldsymbol{\xi}}, \tilde{\boldsymbol{\xi}}^{\mathrm{ind}}) &\leq d_{\mathcal{W}}(\tilde{\boldsymbol{\mu}}_0^{\mathrm{ind}}, \tilde{\boldsymbol{\mu}}_0^{\mathrm{co}}), \end{aligned}$$

where $\tilde{\boldsymbol{\mu}}^{\mathrm{co}}$ is the comonotonic CRV in the same Fréchet class of $\tilde{\boldsymbol{\mu}}^{\mathrm{ind}}$ and $\tilde{\boldsymbol{\mu}}_0^{\mathrm{ind}}$ is the CRV with independent marginals in the same Fréchet class of $\tilde{\boldsymbol{\mu}}_0^{co}$.

When the marginals are generalized gamma CRMs, the specification of the previous bounds together with Theorem 30 brings to the following. In particular, this covers the case where the marginals are gamma random measures, as in (Lijoi et al., 2014; Lijoi & Nipoti, 2014).

Corollary 34. *Let $\tilde{\boldsymbol{\xi}}$ be a* GM*–dependent CRV with marginal generalized gamma random measures with parameters b, σ and total measure α. Then*

$$d_{\mathcal{W}}(\tilde{\boldsymbol{\xi}}, \tilde{\boldsymbol{\xi}}^{\mathrm{co}})^2 \leq \gamma_{b,\sigma}\, \alpha\, z, \qquad\qquad d_{\mathcal{W}}(\tilde{\boldsymbol{\xi}}, \tilde{\boldsymbol{\xi}}^{\mathrm{ind}})^2 \leq \gamma_{b,\sigma}\, \alpha\, (1 - z).$$

where $\gamma_{b,\sigma}$ is the constant defined in (3.18).

As one could expect from the construction in Definition 11, the larger the parameter z, the closer one is to the situation of independence and the farther from the one of exchangeability. Our techniques allow for the derivation of convergence rates for the approximation of exchangeability as $z \to 1$, in terms of the Wasserstein distance.
Figure 3.8 below shows the comparison between the simulated Wasserstein distance and our theoretical upper bound, as z increases, when the marginals are gamma CRMs $(\sigma = 0, b = 1)$.

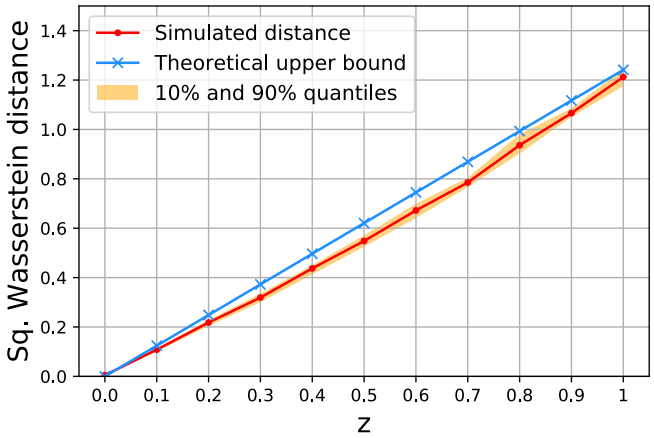

Figure 3.8: Simulation of the Wasserstein distance between a GM–dependent CRV $(\mu_1(\mathbb{X}), \mu_2(\mathbb{X}))$ of parameter z with gamma marginals of shape = scale = 1 and a bivariate distribution with a.s. equal gamma marginals of shape = scale = 1. The simulations were performed with samples of 10000 observations.

In this section we have found tight upper bounds for the distance $d_{\mathcal{W}}$ from comonotonicity for notable homogeneous CRVs, so to exploit the results in Remark 7. Since the Lévy measure factorizes, the supremum of the Wasserstein distance over all Borel sets is always attained on the entire sample space $\mathbb{X}$. Though is it more common to find homogenous CRVs as priors for Bayesian nonparametric models, it should be underlined that finding the supremum could be considerably more complex for non homogeneous CRVs.

3.8 Measuring dependence between random hazards

Most common specifications of the de Finetti measure for exchangeable sequences are expressed as suitable transformations of CRMs. Popular methods include the normalization of Regazzini et al. (2003), which leads to a direct prior on the random probability measure, the exponential transformation of Doksum (1974), which specifies a nonparametric prior for the random survival function, and the tail distribution of Hjort (1990), which brings to a prior for the random cumulative hazards. In this section we focus on

models for almost surely continuous hazards, which are of great interest in the context of survival analysis and reliability theory. If F is an absolutely continuous cumulative distribution function on $[0, +\infty)$, we recall that the hazards are defined as $h = F'/(1-F)$ and represent the instantaneous risk of failure. Random hazards are often modeled as kernel mixtures $\tilde{h}(t) = \int_{\mathbb{X}} k(t|x)\, d\tilde{\mu}(x)$, where $k : \mathbb{R}^+ \times \mathbb{X} \to [0, +\infty)$ is a measurable kernel function and $\tilde{\mu}$ is a CRM with Poisson random measure $\mathcal{N}$ that satisfies

$$\lim_{t\to\infty} \int_0^t \int_{\mathbb{R}^+\times\mathbb{X}} k(u \,|\, y)\, s\, du\, \mathcal{N}(ds, dy) = +\infty. \tag{3.21}$$

Condition (3.21) ensures that the mean cumulative hazards go to $+\infty$ as time increases. This model was initially proposed with a gamma random measure and a specific kernel by Dykstra & Laud (1981), and has been further generalized to generic kernels (Lo & Weng, 1989) and to generic CRMs (James, 2005). If $\tilde{\boldsymbol{\mu}}$ is a random vector of measures,

$$\tilde{\boldsymbol{h}}(t) = \int_{\mathbb{X}} k(t|x)\, \tilde{\boldsymbol{\mu}}(dx) \tag{3.22}$$

defines dependent hazards, which may be used as de Finetti priors for partially exchangeable sequences. Notable examples include hierarchical dependent structures Camerlenghi et al. (2020) and GM-dependent structures Lijoi & Nipoti (2014). The results of Section 3.5 and Section 3.7 may be adapted to quantify the dependence between the random hazards when $\tilde{\boldsymbol{\mu}}$ is a CRV. This brings to a direct measure of dependence between the de Finetti priors corresponding to different groups.

A first key result is Lemma 35 applied to the function $f(\cdot) = k(t|\cdot)$, which brings to the expression $\tilde{\boldsymbol{h}}(t) \overset{\mathrm{d}}{=} \tilde{\boldsymbol{\mu}}_t(\mathbb{X})$ for an appropriate CRV $\tilde{\boldsymbol{\mu}}_t$. Given two measure spaces $\mathbb{X}_1$ and $\mathbb{X}_2$, we recall that if ν is a measure on $\mathbb{X}_1$ and $g : \mathbb{X}_1 \to \mathbb{X}_2$ is a measurable function, the pushforward measure $g\#\nu$ on $\mathbb{X}_2$ is defined by $(g\#\nu)(A) = \nu(g^{-1}(A))$.

Lemma 35. *Let $\tilde{\boldsymbol{\mu}}$ be a CRV with intensity measure ν and let $f : \mathbb{X} \to \mathbb{R}^+$ be a measurable function. Then the random vector of measures $\tilde{\boldsymbol{\mu}}_f(dx) = f(x)\tilde{\boldsymbol{\mu}}(dx)$ is a CRV with Lévy intensity equal to the pushforward measure $\nu_f = p_f \,\#\, \nu$ where $p_f(s_1, s_2, x) = (s_1 f(x), s_2 f(x), x)$.*

Lemma 35 may be seen as a multivariate extension of Lemma 10 in Section 2.4. In particular, we observe that the hazard rates $\tilde{\boldsymbol{h}}^{\mathrm{co}}$ induced by a comonotonic CRV $\tilde{\boldsymbol{\mu}}^{\mathrm{co}}$ through (3.22) are comonotonic, i.e. $\tilde{h}_1^{\mathrm{co}}(t) = \tilde{h}_2^{\mathrm{co}}(t)$ a.s. for every t. Similarly, when $\tilde{\boldsymbol{\mu}}^{\mathrm{ind}}$ is the independent CRV, the induced hazards $\tilde{\boldsymbol{h}}^{\mathrm{ind}}$ are independent. We use this observation to study the Wasserstein distance between the dependent hazards and the two extreme cases of comonotonicity and independence. Corollary 36 deals with the GM–dependent hazards of Lijoi & Nipoti (2014) when the marginals are gamma random measures and the kernel of the type of Dykstra & Laud (1981), $k(t|x) = \beta(y)\mathbb{1}_{[0,t]}(x)$, which is a popular choice for modeling increasing hazards. For simplicity we restrict to constant functions $\beta(s) = \beta$, which are the most common choice in applications. In such scenario one usually considers the base measure of the gamma random measure to be equal to the Lebesgue measure on a large time interval $[0, T]$, i.e. $\alpha P_0(ds) = \mathbb{1}_{[0,T]}(s)\, ds$, so that $\mathbb{X} = \mathbb{R}$.

Corollary 36. *Let $\tilde{\boldsymbol{h}}$ be dependent hazards as defined in (3.22) s.t. $\tilde{\boldsymbol{\mu}}$ is a* GM*–dependent CRV (3.20) with marginal gamma CRM of base measure $\alpha P_0(ds) = \mathbb{1}_{[0,T]}(s)\,ds$ and $k(t|x) = \beta \mathbb{1}_{[0,t]}(x)$, with $\beta > 0$. If $\tilde{\boldsymbol{h}}^{\text{co}}, \tilde{\boldsymbol{h}}^{\text{ind}}$ are in the same Fréchet class as $\tilde{\boldsymbol{h}}$, for every $t \in [0, T]$,*

$$\mathcal{W}(\tilde{\boldsymbol{h}}(t), \tilde{\boldsymbol{h}}^{\text{co}}(t))^2 \leq \gamma_\beta\, t\, z; \qquad \mathcal{W}(\tilde{\boldsymbol{h}}(t), \tilde{\boldsymbol{h}}^{\text{ind}}(t))^2 \leq \gamma_\beta\, t\, (1-z);$$

where $\gamma_\beta = \gamma_{\beta^{-1},0}$ defined in (3.18).

3.9 Proofs

3.9.1 Background results

We first recall some key results concerning the Wasserstein distance. See (Bickel & Freedman, 1981, Lemma 8.6 and 8.8). If $(\boldsymbol{X}^1, \ldots, \boldsymbol{X}^n)$ and $(\boldsymbol{Y}^1, \cdots, \boldsymbol{Y}^n)$ are tuples of independent random vectors on $\mathbb{R}^2$, then

$$\mathcal{W}(\boldsymbol{X}^1 + \cdots + \boldsymbol{X}^n, \boldsymbol{Y}^1 + \cdots + \boldsymbol{Y}^n) \leq \sum_{i=1}^{n} \mathcal{W}(\boldsymbol{X}^i, \boldsymbol{Y}^i). \tag{3.23}$$

Moreover, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are two random vectors on $\mathbb{R}^2$ with finite second moment, then

$$\mathcal{W}(\boldsymbol{X}, \boldsymbol{Y})^2 = \mathcal{W}(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}), \boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y}))^2 + \|\mathbb{E}(\boldsymbol{X}) - \mathbb{E}(\boldsymbol{Y})\|^2. \tag{3.24}$$

Next, if P_1, P_2, Q_1, Q_2 are probability measures, then for every $\alpha \in [0, 1]$

$$\begin{aligned}\mathcal{W}(\alpha\, P_1 + (1-\alpha)\, P_2, \alpha\, Q_1 + (1-\alpha)\, Q_2) \leq \\ \alpha\, \mathcal{W}(P_1, Q_1) + (1-\alpha)\, \mathcal{W}(P_2, Q_2).\end{aligned} \tag{3.25}$$

Furthermore, we recall (Rüschendorf, 1991, Theorem 12) to establish the optimality of a transport map.

Theorem 37 (Rüschendorff 1991)**.** *If $\boldsymbol{X}$ is a random object on $\mathbb{R}^2$ and $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ is continuously differentiable, then $(\boldsymbol{X}, \phi(\boldsymbol{X}))$ is an optimal coupling with respect to the 2–Wasserstein distance if and only if the following hold:*

1. *ϕ is monotone, i.e. $\langle \boldsymbol{x} - \boldsymbol{y}, \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) \rangle \geq 0$ for every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2$, where $\langle \cdot \rangle$ indicates the standard scalar product on $\mathbb{R}^2$;*
2. *The matrix $D\phi = \left(\frac{\partial \phi_i}{\partial x_j}\right)_{i,j}$ is symmetric.*

3.9.2 Proof of Theorem 20

The proof of Theorem 20 is based on the following result, which will also be instrumental to further proofs. As before, F_X denotes the cdf of X.

Theorem 38. *Let X_1, X_2, X be possibly dependent random variables whose law is a.c. w.r.t. the Lebesgue measure on $\mathbb{R}$. Then, for every continuously differentiable $g : \mathbb{R}^2 \to \mathbb{R}$, the map*

$$(x_1, x_2) \mapsto \phi_g(x_1, x_2) = (F_X^{-1} \circ F_{g(X_1,X_2)} \circ g\,(x_1, x_2), F_X^{-1} \circ F_{g(X_1,X_2)} \circ g\,(x_1, x_2)),$$

provides a transportation map between $\mathcal{L}(X_1, X_2)$ and $\mathcal{L}(X, X)$. Moreover,

$$(x_1, x_2) \mapsto \phi(x_1, x_2) = (F_X^{-1} \circ F_{X_1+X_2}(x_1 + x_2), F_X^{-1} \circ F_{X_1+X_2}(x_1 + x_2)),$$

is an optimal transport map.

Proof. First observe that $F_{g(X_1,X_2)} \circ g(X_1, X_2) \sim \mathrm{Unif}([0, 1])$. Since X is a.c. w.r.t. the Lebesgue measure on $\mathbb{R}$, $F_X^{-1} \circ F_{g(X_1,X_2)} \circ g(X_1 + X_2) \overset{\mathrm{d}}{=} X$. This ensures that ϕ_g is indeed a coupling between (X_1, X_2) and (X, X). In order to prove that ϕ is an optimal transport map, we refer to the sufficient conditions described in Theorem 37. Note that

$$\begin{aligned}&\langle \boldsymbol{x} - \boldsymbol{y}, \phi(\boldsymbol{x}) - \phi(\boldsymbol{y})\rangle = \\ &= (x_1 - y_1 + x_2 - y_2)\,(F_X^{-1} \circ F_{X_1+X_2}(x_1 + x_2) - F_X^{-1} \circ F_{X_1+X_2}(y_1 + y_2)).\end{aligned}$$

Since cdfs are non–decreasing functions, and the inverse of a non–decreasing function is non–decreasing as well, F_X^{-1} is non–decreasing. Thus $x_1 + x_2 \leq y_1 + y_2$ if and only if $F_X^{-1} \circ F_{X_1+X_2}(x_1 + x_2) \leq F_X^{-1} \circ F_{X_1+X_2}(y_1 + y_2)$. It follows that the previous expression is always non–negative, and the monotonicity condition holds. As for the symmetry, this easily holds since the two components of ϕ are the same and are symmetric in the two arguments. □

Now consider $\tilde{\boldsymbol{\mu}}(A) = (X_1, X_2)$ and $\tilde{\boldsymbol{\mu}}^{\mathrm{co}}(A) = (X, X)$. Theorem 38 guarantees that

$$\mathcal{W}(\tilde{\boldsymbol{\mu}}(A), \tilde{\boldsymbol{\mu}}^{\mathrm{co}}(A))^2 = \sum_{i=1}^{2} \mathbb{E}\big(\big|\tilde{\mu}_i(A) - F^{-1}_{\tilde{\mu}_1(A)}(F_{\tilde{\mu}_1(A)+\tilde{\mu}_2(A)}(\tilde{\mu}_1(A) + \tilde{\mu}_2(A))\big|\big)^2$$

and note that $F^{-1}_{\tilde{\mu}_1(A)}(F_{\tilde{\mu}_1(A)+\tilde{\mu}_2(A)}(\tilde{\mu}_1(A) + \tilde{\mu}_2(A)) \overset{\mathrm{d}}{=} \tilde{\mu}_1(A)$. Thus, we have

$$\mathcal{W}(\tilde{\boldsymbol{\mu}}(A), \tilde{\boldsymbol{\mu}}^{\mathrm{co}}(A))^2 = 4\,(\mathbb{E}\big(\tilde{\mu}_1(A)^2\big) - \omega_{\tilde{\boldsymbol{\mu}},A}).$$

3.9.3 Proof of Lemma 21

Let $\tilde{\boldsymbol{\mu}}(A) = (X_1, X_2)$, so that $\omega_{\tilde{\boldsymbol{\mu}},A} = \mathbb{E}(X_1 F_{X_1}^{-1}(F_{X_1+X_2}(X_1 + X_2)))$. Since $\mathcal{L}(X_1, X_2) = \mathcal{L}(X_2, X_1)$,

$$\mathbb{E}(X_1 F_{X_1}^{-1}(F_{X_1+X_2}(X_1 + X_2))) = \frac{1}{2}\mathbb{E}((X_1 + X_2) F_{X_1}^{-1}(F_{X_1+X_2}(X_1 + X_2))).$$

We conclude with a change of variable $U = F_{X_1+X_2}(X_1 + X_2) \sim \mathrm{Unif}([0, 1])$.

3.9.4 Proof of Corollary 22

The proof is based on Theorem 20. First observe that $\tilde{\mu}_1(A) \sim \text{gamma}(\alpha\, P_0(A))$. Thus $\mathbb{E}(\tilde{\mu}_1(A)^2) = \alpha P_0(A)(1+\alpha P_0(A))$. Moreover, $\omega_{\tilde{\boldsymbol{\mu}},A}$ can be rewritten as

$$\mathbb{E}\big(\tilde{\mu}_1(A)\, S^{-1}_{\tilde{\mu}_1(A)}(S_{\tilde{\mu}_1(A)+\tilde{\mu}_2(A)}(\tilde{\mu}_1(A)+\tilde{\mu}_2(A))\big),$$

where S_X denotes the survival function. Now, since $\tilde{\mu}_1(A)$ and $\tilde{\mu}_2(A)$ are independent, $\tilde{\mu}_1(A)+\tilde{\mu}_2(A) \sim \text{gamma}(2\,\alpha\, P_0(A))$. Thus, we have

$$\omega_{\tilde{\boldsymbol{\mu}},A} = \int_0^{+\infty}\int_0^{+\infty} s_1\, \text{Inv}\Gamma_{\alpha\, P_0(A)}\left(\frac{\Gamma(\alpha\, P_0(A))}{\Gamma(2\,\alpha\, P_0(A))}\Gamma(2\,\alpha\, P_0(A), s_1+s_2)\right)\cdot \\ \cdot\, \rho_{\alpha\, P_0(A)}(s_1)\, \rho_{\alpha\, P_0(A)}(s_2)\, ds_1\, ds_2$$

with ρ_ϕ the density function of a gamma(ϕ,1). With a change of variables $(t_1,t_2) = (s_1, s_1+s_2)$ this is equal to

$$\int_0^{+\infty} \text{Inv}\Gamma_{\alpha\, P_0(A)}\left(\frac{\Gamma(\alpha\, P_0(A))}{\Gamma(2\,\alpha\, P_0(A))}\Gamma(2\,\alpha\, P_0(A), t_2)\right)\cdot \\ \cdot \int_0^{t_2} t_1\, \rho_{\alpha\, P_0(A)}(t_1)\, \rho_{\alpha\, P_0(A)}(t_2-t_1)\, dt_1\, dt_2.$$

Now, $t_1\, \rho_{\alpha\, P_0(A)}(t_1) = \alpha\, P_0(A)\, \rho_{\alpha\, P_0(A)+1}(t_1)$, so that

$$\int_0^{t_2} t_1\, \rho_{\alpha\, P_0(A)}(t_1)\, \rho_{\alpha\, P_0(A)}(t_2-t_1)\, dt_1\, dt_2$$

is proportional to the convolution between two gamma random variables with parameters, respectively, ($\alpha\, P_0(A)+1$,1) and ($\alpha\, P_0(A)$,1), evaluated in t_2. This corresponds to the density of a gamma($2\,\alpha\, P_0(A)+1$,1) random variable evaluated in t_2. Thus $\omega_{\tilde{\boldsymbol{\mu}},A}$ is equal to

$$\frac{\alpha\, P_0(A)}{\Gamma(2\,\alpha\, P_0(A)+1)}\int_0^{+\infty} \text{Inv}\Gamma_{\alpha\, P_0(A)}\left(\frac{\Gamma(\alpha\, P_0(A))}{\Gamma(2\,\alpha\, P_0(A))}\Gamma(2\,\alpha\, P_0(A), t)\right) e^{-t}\, t^{2\,\alpha\, P_0(A)}\, dt.$$

3.9.5 Proof of Theorem 23

We show that for every real sequence $\{r_n \,|\, n \in \mathbb{N}\}$ s.t. $\lim_{n\to+\infty} r_n = +\infty$,

$$\mathcal{W}(\tilde{\boldsymbol{\mu}}^1(A), \tilde{\boldsymbol{\mu}}^2(A)) \leq \lim_{n\to+\infty} \sqrt{r_n}\, \mathcal{W}(\rho^1_{r_n,A,D_1}, \rho^2_{r_n,A,D_2}). \tag{3.26}$$

Since both complementary families D_1, D_2 have continuously increasing mass, there exists n_0 s.t. for every $n > n_0$ there exist $\epsilon^1_{n,A}, \epsilon^2_{n,A} > 0$ s.t.

$$r_n = \nu^1(D(\epsilon^1_{n,A}) \times A) = \nu^2(D(\epsilon^2_{n,A}) \times A). \tag{3.27}$$

Before moving to the core of the proof, we show that

$$\lim_{n\to+\infty} \epsilon^i_{n,A} = 0. \tag{3.28}$$

We reason by contradiction. Supposing (3.28) does not hold, there must be a subsequence $\{\epsilon^i_{h_n,A}\}$ converging to a (possibly infinite) limit $\epsilon^i_* \neq 0$. Since $\lim_{n\to+\infty} r_n = +\infty$, also $\lim_{n\to+\infty} r_{h_n} = +\infty$. Then there is at least one increasing subsequence $\{r_{k_n} \,|\, n \in \mathbb{N}\} \subset \{r_{h_n} \,|\, n \in \mathbb{N}\}$ s.t. $\lim_{n\to+\infty} \epsilon^i_{k_n,A} = \epsilon^i_*$ and $\lim_{n\to+\infty} r_{k_n} = +\infty$.
Since D is increasing and ν is monotone, $r_{k_n} \leq r_{k_{n+1}}$ implies $D(\epsilon^i_{k_n,A}) \subset D(\epsilon^i_{k_{n+1},A})$. Thus by the monotone convergence theorem,

$$+\infty = \lim_{n\to+\infty} \nu^i(D(\epsilon^i_{k_n,A}) \times A) = \nu^i(D(\epsilon^i_*) \times A).$$

Given the Lévy intensity is finite outside of the origin by (3.3), $\nu^i(D(\epsilon^i_*) \times A) < +\infty$, which is a contradiction. Thus (3.28) holds.
Now recall that by (3.2) there exist Poisson random measures $\mathcal{N}^i$ s.t. for every $A \in \mathcal{X}$, $\tilde{\boldsymbol{\mu}}^i(A) = \int_{\mathbb{R}^2_+ \times A} \boldsymbol{s}\, \mathcal{N}^i(ds_1, ds_2, dx)$, for $i = 1, 2$. Since the evaluations of Poisson random measures on disjoint sets are independent, by (3.23) for every $n > 0$,

$$\begin{aligned}
&\mathcal{W}(\tilde{\boldsymbol{\mu}}^1(A), \tilde{\boldsymbol{\mu}}^2(A)) \leq \\
&\mathcal{W}\Big(\int_{B_1(\epsilon^1_{n,A})\times A} \boldsymbol{s}\, \mathcal{N}^1(ds_1, ds_2, dx), \int_{B_2(\epsilon^2_{n,A})\times A} \boldsymbol{s}\, \mathcal{N}^2(ds_1, ds_2, dx) \Big) && (3.29)\\
&+ \mathcal{W}\Big(\int_{D_1(\epsilon^1_{n,A})\times A} \boldsymbol{s}\, \mathcal{N}^1(ds_1, ds_2, dx), \int_{D_2(\epsilon^2_{n,A})\times A} \boldsymbol{s}\, \mathcal{N}^2(ds_1, ds_2, dx) \Big). && (3.30)
\end{aligned}$$

We prove that the first summand (3.29) goes to zero as $n \to +\infty$. By bounding the Wasserstein distance with the second moments as in (3.5) and using the properties of Poisson random measures, (3.29) is bounded from above by

$$\left(2 \sum_{\substack{i=1,2\\ j=1,2}} \int_{B_i(\epsilon^i_{n,A})} s_j^2\, \nu^i(ds_1, ds_2, A) + \Big(\int_{B_i(\epsilon^i_{n,A})} s_j\, \nu^i(ds_1, ds_2, A) \Big)^2 \right)^{\frac{1}{2}}$$

Thanks to the finiteness of the integrals in (3.6) and (3.7), we may apply the dominated convergence theorem and bring the limit as $n \to +\infty$ inside both integrals. In order to prove that the above expression goes to zero we thus need to show that

$$\int_{\mathbb{R}^2_+} \mathbb{1}_{\cap_{n\in\mathbb{N}} B_i(\epsilon^i_{n,A})}(s_1, s_2)\, s_j^k\, \nu^i(ds_1, ds_2, A) = 0,$$

where $i, j, k = 1, 2$. By absolute continuity of the integral it suffices to show that $\nu^i(\cap_{n\in\mathbb{N}} B_i(\epsilon^i_{n,A}) \times A) = 0$. Now, by assumptions on the family B, we know that $\nu^i(\cap_{\epsilon\in(0,1]} B_i(\epsilon) \times A) = 0$. We then prove that

$$\nu^i(\cap_{n\in\mathbb{N}} B_i(\epsilon^i_{n,A}) \times A) \leq \nu^i(\cap_{\epsilon\in(0,1]} B_i(\epsilon) \times A) = 0, \tag{3.31}$$

by showing that $\cap_{n\in\mathbb{N}} B_i(\epsilon^i_{n,A}) \subset \cap_{\epsilon\in(0,1]} B_i(\epsilon)$. Let $x \in \cap_{n\in\mathbb{N}} B_i(\epsilon^i_{n,A})$. Since $\lim_{n\to+\infty} \epsilon^i_{n,A} = 0$ by (3.28), for every $\epsilon \in (0,1]$ there exists n s.t. $\epsilon^i_{n,A} < \epsilon$. Since B_i is an increasing family, $x \in B_i(\epsilon^i_{n,A}) \subset B_i(\epsilon)$. Thus $x \in \cap_{\epsilon>0} B_i(\epsilon)$.
As for the second summand (3.30), since the Lévy intensities are bounded outside of the origin by (3.3), $\mathbb{1}_{D_i(\epsilon^i_{n,A})}(\boldsymbol{s})\,\mathcal{N}^i(ds_1, ds_2, dy)$ is a Poisson random measure with finite mean for $i = 1, 2$. Thus by (Sato, 1999, Proposition 19.5) their integrals have a compound Poisson distribution on $\mathbb{R}^2$ with intensity measure $\int_A \mathbb{1}_{D_i(\epsilon^i_{n,A})}(\boldsymbol{s})\,\nu^i(ds_1, ds_2, dy)$ and same total measure r_n. Hence we have

$$\int_{D_i(\epsilon^i_{n,A})\times A} \boldsymbol{s}\, \mathcal{N}^i(ds_1, ds_2, dx) \overset{\mathrm{d}}{=} \sum_{j=1}^{N^i} \boldsymbol{X}^i_j,$$

where N^i has a Poisson distribution with mean r_n and is independent of $\{\boldsymbol{X}^i_j \,|\, j \geq 1\}$, which are iid random variables with distribution $\rho^i_{r_n,A,D_i}$. Proposition 24 thus entails

$$\mathcal{W}\left(\sum_{j=1}^{N^1} \boldsymbol{X}^1_j, \sum_{j=1}^{N^2} \boldsymbol{X}^2_j\right) \leq \sqrt{r_n}\, \mathcal{W}(\rho^1_{r_n,A,D_1}, \rho^2_{r_n,A,D_2}) + (r_n^2 + r_n)\, \|\mathbb{E}(\boldsymbol{X}^1_1) - \mathbb{E}(\boldsymbol{X}^2_1)\|^2.$$

Now, $(r_n^2 + r_n)\, \|\mathbb{E}(\boldsymbol{X}^1_1) - \mathbb{E}(\boldsymbol{X}^2_1)\|^2$ is equal to

$$\left(1 + \frac{1}{r_n}\right) \sum_{i=1,2} \left| \int_{D_1(\epsilon^1_{n,A})} s_i\, \nu^1(ds_1, ds_2, A) - \int_{D_2(\epsilon^2_{n,A})} s_i\, \nu^2(ds_1, ds_2, A) \right|^2,$$

which as $n \to +\infty$ by the monotone convergence theorem converges to

$$\sum_{i=1,2} \left| \int_{\mathbb{R}^2_+} s_i\, \nu^1(ds_1, ds_2, A) - \int_{\mathbb{R}^2_+} s_i\, \nu^2(ds_1, ds_2, A) \right|^2 \tag{3.32}$$

Since the vectors are in the same Fréchet class, (3.32) is equal to 0. The bound in (3.26) hence follows by taking the limit as n goes to $+\infty$. In order to prove its finiteness it suffices to observe that by (3.5), $\sqrt{r_n}\, \mathcal{W}(\rho^1_{r_n,A,D_1}, \rho^2_{r_n,A,D_2})$ is bounded from above by the square root of

$$2 \sum_{i=1}^{2} \int_{\mathbb{R}^2_+} (s_1^2 + s_2^2)\, \nu^i(ds_1, ds_2, A) + \left(\int_{\mathbb{R}^2_+} (s_1 + s_2)\, \nu^i(ds_1, ds_2, A) \right)^2,$$

which is finite by (3.6) and (3.7).
We now show that the limit as n goes to $+\infty$ does not depend on the choice of compatible families D_1 and D_2. First we prove that given a bivariate Lévy intensity ν with compatible families D and D^*,

$$\lim_{n\to+\infty} \sqrt{r_n}\, \mathcal{W}(\rho_{r_n,A,D}, \rho_{r_n,A,D^*}) = 0. \tag{3.33}$$

For every n consider $\epsilon_{n,A}$ and $\epsilon^*_{n,A}$ as in (3.27). Let then $\Omega(n) = D(\epsilon_{n,A}) \cap D^*(\epsilon^*_{n,A})$ and denote by $q_n = \nu(\Omega(n) \times A)$. We define

$$P^0_n(ds_1, ds_2) = \frac{1}{q_n} \mathbb{1}_{\Omega(n)}(s_1, s_2)\, \nu(ds_1, ds_1, A)$$
$$P_n(ds_1, ds_2) = \frac{1}{r_n - q_n} \mathbb{1}_{D(\epsilon_{n,A})\setminus\Omega(n)}(s_1, s_2)\, \nu(ds_1, ds_1, A)$$
$$P^*_n(ds_1, ds_2) = \frac{1}{r_n - q_n} \mathbb{1}_{D^*(\epsilon^*_{n,A})\setminus\Omega(n)}(s_1, s_2)\, \nu(ds_1, ds_1, A)$$

and consider the decompositions

$$\rho_{r_n,A,D} = \frac{q_n}{r_n} P^0_n + \frac{r_n - q_n}{r_n} P_n \qquad \rho_{r_n,A,D^*} = \frac{q_n}{r_n} P^0_n + \frac{r_n - q_n}{r_n} P^*_n.$$

By the convexity property in (3.25), since P^0_n is a shared component,

$$\mathcal{W}(\rho_{r_n,A,D}, \rho_{r_n,A,D^*}) \le \frac{r_n - q_n}{r_n} \mathcal{W}(P_n, P^*_n).$$

Hence by (3.5), $\sqrt{r_n}\, \mathcal{W}(\rho_{r_n,A,D}, \rho_{r_n,A,D^*})$ is bounded from above by the squared root of

$$\frac{r_n - q_n}{r_n} 4 \left(\int_{\mathbb{R}^2_+} (s_1^2 + s_2^2)\, \nu(ds_1, ds_2, A) + \left(\int_{\mathbb{R}^2_+} (s_1 + s_2)\, \nu(ds_1, ds_2, A) \right)^2 \right),$$

Since $D(\epsilon_{n,A}) \setminus \Omega(n) \subset D^*(\epsilon^*_{n,A})^c$, $r_n - q_n = \nu(D(\epsilon_{n,A}) \setminus \Omega(n) \times A) \le \nu(D^*(\epsilon^*_{n,A})^c \times A)$. Thus, by reasoning as in (3.31),

$$\limsup_{n\to+\infty} r_n - q_n \le \limsup_{n\to+\infty} \nu(D^*(\epsilon^*_{n,A})^c \times A) = \nu(\cap_{n\in\mathbb{N}} B^*(\epsilon^*_{n,A}) \times A) = 0.$$

Hence, $\lim_{n\to+\infty} r_n - q_n = 0$ and we conclude that $\lim_{r_n\to+\infty} \sqrt{r_n}\mathcal{W}(\rho_{r_n,A,D}, \rho_{r_n,A,D^*}) = 0$.

Now, consider two compatible families D^*_1 and D^*_2 for ν^1 and ν^2, respectively. By the triangular inequality, $\mathcal{W}(\rho^1_{r_n,A,D^*_1}, \rho^2_{r_n,A,D^*_2})$ is bounded from above by

$$\mathcal{W}(\rho^1_{r_n,A,D^*_1}, \rho^1_{r_n,A,D_1}) + \mathcal{W}(\rho^1_{r_n,A,D_1}, \rho^2_{r_n,A,D_2}) + \mathcal{W}(\rho^2_{r_n,A,D_2}, \rho^2_{r_n,A,D^*_2})$$

Then, thanks to (3.33) by taking the limit as $n \to +\infty$,

$$\lim_{n\to+\infty} \sqrt{r_n}\, \mathcal{W}(\rho^1_{r_n,A,D^*_1}, \rho^2_{r_n,A,D^*_2}) \le \lim_{n\to+\infty} \sqrt{r_n}\, \mathcal{W}(\rho^1_{r_n,A,D_1}, \rho^2_{r_n,A,D_2})$$

Equality follows by changing the role of (D_1, D_2) and (D^*_1, D^*_2) in the previous argument.

3.9.6 Proof of Proposition 24

We rely on the key identity (3.24). First we observe that $\mathbb{E}(\boldsymbol{X}) = r\,\mathbb{E}(\boldsymbol{X}_1)$ and $\mathbb{E}(\boldsymbol{Y}) = r\,\mathbb{E}(\boldsymbol{Y}_1)$. By considering the couplings s.t. $N_x = N_Y$ a.s.,

$$\begin{aligned}
&\mathcal{W}(\boldsymbol{X} - r\,\mathbb{E}(\boldsymbol{X}^1), \boldsymbol{Y} - r\,\mathbb{E}(\boldsymbol{Y}^1))^2 \\
&\leq \inf_{C((\boldsymbol{X}^i)_{i\geq 1},(\boldsymbol{Y}^i)_{i\geq 1})} \mathbb{E}\left(\left\| \sum_{i=1}^{N_x} \boldsymbol{X}^i - r\,\mathbb{E}(\boldsymbol{X}^1) - \sum_{i=1}^{N_x} \boldsymbol{Y}^i + r\,\mathbb{E}(\boldsymbol{Y}^1)\right\|^2\right) \\
&= \inf_{C((\boldsymbol{X}^i)_{i\geq 1},(\boldsymbol{Y}^i)_{i\geq 1})} \mathbb{E}\left(\mathrm{Var}\left(\sum_{i=1}^{N_x} X_1^i - r\,\mathbb{E}(X_1^1) - \sum_{i=1}^{N_x} Y_1^i - r\,\mathbb{E}(Y_1^1)\,\middle|\, N_x\right)\right) + \\
&\qquad + \mathbb{E}\left(\mathrm{Var}\left(\sum_{i=1}^{N_x} X_2^i - r\,\mathbb{E}(X_2^1) - \sum_{i=1}^{N_x} Y_2^i - r\,\mathbb{E}(Y_2^1)\,\middle|\, N_x\right)\right)
\end{aligned}$$

By considering couplings s.t. $(\boldsymbol{X}^i - \boldsymbol{Y}^i)_{i\geq 1}$ are independent and identically distributed,

$$\begin{aligned}
&\leq \inf_{C(\boldsymbol{X}^1,\boldsymbol{Y}^1)} \mathbb{E}(N_x\,\mathrm{Var}(X_1^1 - Y_1^1)) + \mathbb{E}(N_x\,\mathrm{Var}(X_2^1 - Y_2^1)) \\
&= r\,\mathcal{W}(\boldsymbol{X}^1, \boldsymbol{X}^2)^2 - r\,\|\mathbb{E}(\boldsymbol{X}^1 - \boldsymbol{Y}^1)\|^2.
\end{aligned}$$

Finally, by applying (3.24) we conclude the proof.

3.9.7 Proof of Proposition 25

A CRV $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \tilde{\mu}_2)$ is comonotonic if $\tilde{\mu}_1 = \tilde{\mu}_2$ a.s. By uniqueness of the Lévy intensity is suffices to show that ν^{co} induces exchangeability. Consider the set $D = \{(s_1, s_2)\,|\,(s_1, s_2) \in \mathbb{R}_+^2,\, s_1 \neq s_2\}$. By definition of Poisson random measure, for every $A \in \mathcal{X}$

$$\mathbb{P}(\mathcal{N}(D \times A) = 0) = \exp\{-\nu^{\mathrm{co}}(D \times A)\} = 1.$$

Thus with probability 1,

$$\tilde{\mu}_1(A) = \int_{\mathbb{R}_+^2 \times A} s_1\,\mathcal{N}(ds_1, ds_2, dx) = \int_{\mathbb{R}_+^2 \times A} s_2\,\mathcal{N}(ds_1, ds_2, dx) = \tilde{\mu}_2(A).$$

3.9.8 Proof of Theorem 26

Let $(X_1, X_2) \sim \rho_{r,A,D}$ and $(X, X) \sim \rho^{\mathrm{co}}_{r,A,D^{\mathrm{co}}}$. For every continuously differentiable function g, we define

$$K^g_{r,\nu,A,D,D^{\mathrm{co}}} = \sum_{i=1}^{2} \int_{\mathbb{R}_+^2} |s_i - F_X^{-1} \circ F_{g(X_1,X_2)} \circ g\ (s_1, s_2)|^2\,\rho_{r,A}(ds_1\,ds_2).$$

Theorem 38 guarantees that $\mathcal{W}(\rho_{r,A,D}, \rho^{\mathrm{co}}_{r,A,D^{\mathrm{co}}})^2 \leq K^g_{r,\nu,A,D,D^{\mathrm{co}}}$, and the equality holds for $g(s_1, s_2) = s_1 + s_2$. In order to find the limit of $r\,K^g_{r,\nu,A,D,D^{\mathrm{co}}}$ as $r \to +\infty$, we must

establish the conditions for the monotone convergence theorem. Since by Theorem 23 the limit does not depend on the compatible families D and D^{co}, we choose $D = D^g = (B^g)^c$ defined in (3.10), and $D^{\text{co}} = D^+$ as in (1) of Figure 3.2. First rewrite the bound as

$$\sum_{i=1}^{2} \int_{\mathbb{R}^2_+} |s_i - S_X^{-1} \circ S_{g(X_1,X_2)}(g(s_1,s_2))|^2 \, \mathbb{1}_{(\epsilon_{r,A},+\infty)}(g(s_1,s_2)) \, \nu(ds_1, ds_2, A),$$

where S_X is the survival function of X. The choice $D^{\text{co}} = D^+$ guarantees that $S_X(t) = \frac{1}{r} U_A^\pi(t) \, \mathbb{1}_{(\epsilon/2,+\infty)}(t)$; see Figure 3.3. Thus $\forall s \in (0,1]$, $S_X^{-1}(s) = (U_A^\pi)^{-1}(r\, s)$. On the other hand, $S_{g(X_1,X_2)}(t) = r^{-1} \, h^g_{r,\nu,A}(t)$, where

$$h^g_{r,\nu,A}(t) = \int_{\mathbb{R}^2_+} \mathbb{1}_{(t,+\infty)}(g(t_1,t_2)) \mathbb{1}_{(\epsilon_{r,A},+\infty)}(g(t_1,t_2)) \, \nu(dt_1, dt_2, A).$$

Thus $r\, K^g_{r,\nu,A,D^g,D^+}$ is equal to

$$\sum_{i=1}^{2} \int_{\mathbb{R}^2_+} |s_i - (U_A^\pi)^{-1}(h^g_{r,\nu,A}(g(s_1,s_2)))|^2 \, \mathbb{1}_{(\epsilon_{r,A},+\infty)}(g(s_1,s_2)) \, \nu(ds_1, ds_2, A).$$

Since for every (s_1,s_2) in the domain of integration $g(s_1,s_2) > \epsilon_{r,A}$, every (t_1,t_2) s.t. $g(t_1,t_2) > g(s_1,s_2)$ satisfies $g(t_1,t_2) > \epsilon_{r,A}$. Thus for every (s_1,s_2) in the domain of integration,

$$h^g_{r,\nu,A}(g(s_1,s_2)) = \int_{\mathbb{R}^2_+} \mathbb{1}_{(g(s_1,s_2),+\infty)}(g(t_1,t_2)) \, \nu(dt_1, dt_2, A) = h^g_{\nu,A}(g(s_1,s_2)),$$

where $h^g_{\nu,A}$ is defined in (3.12). The statement in (3.14) follows by the monotone convergence theorem as $r \to +\infty$.

3.9.9 Proof of Theorem 27

We first provide a preliminary result, whose proof we report because it does not seem to be readily available in the literature.

Lemma 39. *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be an integrable non–increasing function and $f^{-1}(x) = \sup\{t \,|\, f(t) \le x\}$ its generalized inverse. Then*

$$\int_0^{+\infty} f(x) \; dx = \int_0^{+\infty} f^{-1}(z) \, dz.$$

Proof. Consider the change of variable $z = f(x)$. Since f is integrable, $\lim_{x\to+\infty} f(x) = 0$. Moreover since f is monotone its derivative is well defined almost everywhere. Thus

$$\int_0^{+\infty} f(x) \, dx = -\int_0^{f(0)} z \frac{1}{f'(f^{-1}(z))} \, dz = -\int_0^{+\infty} z \, (f^{-1})'(z) \, dz$$

having set $f(0) = \lim_{x\to 0^+} f(x) \in [0, +\infty]$. By integration by parts this is equal to

$$= -z\, f^{-1}(z)\,\big|_0^{f(0)} + \int_0^{f(0)} f^{-1}(z)\, dz.$$

If $f(0) < +\infty$, the first summand is clearly 0. Otherwise, we observe that

$$-z\, f^{-1}(z)\,\big|_0^{+\infty} = x\, f(x)\,\big|_0^{+\infty} = 0,$$

because of the integrability assumption. Thus in either case,

$$\int_0^{+\infty} f(x)\, dx = \int_0^{f(0)} f^{-1}(z)\, dz = \int_0^{+\infty} f^{-1}(z)\, dz,$$

since if $f(0) < +\infty$, f^{-1} is equal to zero on the interval $(f(0), +\infty)$. □

Theorem 26 ensures that the limit, as $r \to +\infty$, of $r\,\mathcal{W}(\rho_{r,A,D}, \rho^{\mathrm{co}}_{r,A,D^{\mathrm{co}}})$ is equal to

$$2\int_{\mathbb{R}_+^2} \left| s_1 - (U_A^\pi)^{-1}\Big(\int_{\mathbb{R}_+^2} \mathbb{1}_{(s_1+s_2,+\infty)}(t_1+t_2)\,\nu(t_1,t_2)\,dt_1\,dt_2 \Big)\right|^2 \nu(s_1,s_2,A)\,ds_1\,ds_2.$$

By expanding the square of the binomial, the integral is divided in three summands. We treat them separately. First

$$\int_{\mathbb{R}_+^2} s_1^2\,\nu(s_1,s_2,A)\,ds_1\,ds_2 = \int_{\mathbb{R}_+} s_1^2\,\pi(s_1,A)\,ds_1.$$

Next, with a change of variable $(z_1, z_2) = (s_1, s_1 + s_2)$,

$$\begin{aligned} &\int_{\mathbb{R}_+^2} (U_A^\pi)^{-1}\Big(\int_{\{t_1+t_2>s_1+s_2\}} \nu(t_1,t_2,A)\,dt_1\,dt_2\Big)^2 \nu(s_1,s_2,A)\,ds_1\,ds_2 \qquad (3.34)\\ &= \int_0^{+\infty} (U_A^\pi)^{-1}\Big(\int_{\{t_1+t_2>z_2\}} \nu(t_1,t_2,A)\,dt_1\,dt_2\Big)^2 \Big(\int_0^{z_2} \nu(z_1, z_2-z_1, A)\,dz_1\Big) dz_2 \end{aligned}$$

Simple calculations on the derivative of an integral lead to

$$\frac{d}{dz}\int_{\{t_1+t_2>z\}} \nu(t_1,t_2,A)\,dt_1\,dt_2 = \int_0^z \nu(t_1, z-t_1, A)\,dt_1$$

Thus with a change of variable $s = \int_{\{t_1+t_2>z_2\}} \nu(t_1,t_2,A)\,dt_1\,dt_2$, the integral in (3.34) is equal to $\int_0^{+\infty} (U^\pi)^{-1}(s)^2\,ds$. The function $U_A^\pi(\sqrt{s})$ is non–decreasing and has inverse $|(U_A^\pi)^{-1}(s)|^2$. By applying Lemma 39, its integral on $(0, +\infty)$ is equal to

$$\int_0^{+\infty} U_A^\pi(\sqrt{s})\,ds = \int_0^{+\infty}\int_{\sqrt{s}}^{+\infty} \pi(dt, A)\,ds = \int_0^{+\infty} t^2\,\pi(dt, A)$$

Finally, the expression of the third summand, which is equal to $K_{\nu,A}$ in the statement, derives from the same change of variables.

3.9.10 Proof of Theorem 28

The proof is similar to the one of Theorem 27. By looking at the support of the Lévy intensity in Figure 3.4, Theorem 26 ensures that the limit as $r \to +\infty$ of $r\,\mathcal{W}(\rho_{r,A,D}, \rho^{\mathrm{co}}_{r,A,D^{\mathrm{co}}})$ is equal to

$$2 \int_{\mathbb{R}_+^2} |s_1 - (U_A^\pi)^{-1}(2\,U_A^\pi(s_1+s_2))|^2\, \nu(s_1,s_2,A)\, ds_1\, ds_2.$$

As in the previous proof, the integral is divided in three summands, which we treat separately.

$$\int_{\mathbb{R}_+^2} s_1^2\, \nu(s_1,s_2,A)\, ds_1\, ds_2 = \int_{\mathbb{R}_+} s_1^2\, \pi(s_1,A)\, ds_1.$$

Next, by looking at the support of the Lévy intensity,

$$\begin{aligned}&\int_{\mathbb{R}_+^2} (U_A^\pi)^{-1}(2\,U_A^\pi(s_1+s_2))^2\, \nu(s_1,s_2,A)\, ds_1\, ds_2 \\ &= 2\int_0^\infty (U_A^\pi)^{-1}(2\,U_A^\pi(s))^2\, \pi(s,A)\, ds.\end{aligned}$$

Since $\frac{d}{ds}U_A^\pi(s) = -\pi(s,A)$, with a change of variable $s = 2\,U_A^\pi(s)$, it is equal to $\int_0^{+\infty}(U^\pi)^{-1}(s)^2\, ds$. By reasoning as in Theorem 27, this is equal to $\int_{\mathbb{R}_+} s_1^2\, \pi(s_1,A)\, ds_1$ as well. Finally, since the integrand is equal to zero on the vertical axis, we have

$$\begin{aligned}&\int_{\mathbb{R}_+^2} s_1\, (U_A^\pi)^{-1}(2\,U_A^\pi(s_1+s_2))\, \nu(s_1,s_2,A)\, ds_1\, ds_2 \\ &= \int_0^\infty s_1\, (U_A^\pi)^{-1}(2\,U_A^\pi(s_1))\, \pi(s_1,A)\, ds_1.\end{aligned}$$

3.9.11 Proof of Theorem 31

The proof is based on Theorem 27. Since the Lévy intensities are homogeneous, we apply (3.16). The marginals are gamma random measures of shape parameter 1, thus $U^\pi(t) = E_1(t)$ and

$$\int_0^{+\infty} s^2\, \pi(s)\, ds = \int_0^{+\infty} s\, e^{-s}\, ds = 1.$$

As for the other quantities appearing in (3.16), we observe that if ρ_ϕ is the density of a gamma$(\phi, 1)$ distribution,

$$\begin{aligned}&\int_{\mathbb{R}_+^2} \mathbb{1}_{(t,+\infty)}(z_1+z_2)\, \nu(z_1,z_2)\, dz_1\, dz_2 \\ &= \int_0^1 \left(\int_{\mathbb{R}_+^2} \mathbb{1}_{(t,+\infty)}(z_1+z_2)\, \rho_\phi\left(\frac{z_1}{u}\right) \rho_\phi\left(\frac{z_2}{u}\right) \frac{1}{u^2}\, dz_1\, dz_2 \right) \frac{(1-u)^{\phi-1}}{u}\, du\end{aligned}$$

With a change of variables $(v_1, v_2) = \left(\frac{z_1}{u}, \frac{z_2}{u}\right)$,

$$= \int_0^1 \left(\int_{\mathbb{R}_+^2} \mathbb{1}_{\left(\frac{t}{u}, +\infty\right)} (v_1 + v_2)\, \rho_\phi(v_1)\, \rho_\phi(v_2)\, dv_1\, dv_2 \right) \frac{(1-u)^{\phi-1}}{u}\, du$$
$$= \int_0^1 \mathbb{P}\left\{ X_1 + X_2 > \frac{t}{u} \right\} \frac{(1-u)^{\phi-1}}{u}\, du,$$

where $X_1, X_2 \overset{\text{iid}}{\sim} \text{gamma}(\phi, 1)$ random variables. Thus $X_1 + X_2 \sim \text{gamma}(2\phi, 1)$ and its survival function in $\frac{t}{u}$ is equal to $\frac{\Gamma\left(2\phi, \frac{t}{u}\right)}{\Gamma(2\phi)}$. Next, we observe that

$$\int_0^t s\, \nu(s, t-s)\, ds =$$
$$= \int_0^1 \frac{(1-u)^{\phi-1}}{u^3} \left(\int_0^t s\, \rho_\phi\left(\frac{s}{u}\right) \rho_\phi\left(\frac{t-s}{u}\right) ds \right) du$$

With a change of variable $v = \frac{s}{u}$,

$$= \int_0^1 \frac{(1-u)^{\phi-1}}{u} \left(\int_0^{\frac{t}{u}} v\, \rho_\phi(v)\, \rho_\phi\left(\frac{t}{u} - v\right) ds \right) du.$$

Now, $v\, \rho_\phi(v) = \phi\, \rho_{\phi+1}(v)$. Thus the inner integral is ϕ times the convolution of ρ_ϕ and $\rho_{\phi+1}$ evaluated in $\frac{t}{u}$. Now, if $X \sim \text{gamma}(\phi, 1)$ is independent from $Y \sim \text{gamma}(\phi+1, 1)$, $X + Y \sim \text{gamma}(2\phi + 1, 1)$. Thus

$$\int_0^{\frac{t}{u}} v\, \rho_\phi(v)\, \rho_\phi\left(\frac{t}{u} - v\right) ds = \frac{\phi}{\Gamma(2\phi+1)} e^{-\frac{t}{u}} \left(\frac{t}{u}\right)^{2\phi},$$

from which the final expression for the integral easily follows. We now sketch the proof for ϕ integer:

$$\Gamma\left(2\phi, \frac{t}{u}\right) = (2\phi - 1)! e^{-\frac{t}{u}} \sum_{k=0}^{2\phi-1} \frac{1}{k!} \left(\frac{t}{u}\right)^k.$$

Thus $e(\phi, t)$ is equal to

$$\sum_{k=0}^{2\phi-1} \frac{t^k}{k!} \int_0^1 e^{-\frac{t}{u}} (1-u)^{\phi-1} u^{-k-1} du,$$

which coincides with a sum over k of $f(\phi, k, t)$. In order to derive the expression for $f(\phi, n, t)$ for a generic integer n, we apply the binomial formula

$$(1-u)^{\phi-1} = \sum_{j=0}^{\phi-1} \binom{\phi-1}{j} (-u)^j,$$

from which we easily derive

$$g(n,j,t) = \int_0^1 e^{-\frac{t}{u}} u^{-n-1+j} du = t^{-n+j}\,\Gamma(n-j,t).$$

The final expression derives from writing $\Gamma(k,t)$ as a sum, both when k is a positive integer and when it is a negative one.

3.9.12 Proof of Theorem 32

By resorting to (3.19), one derives the expression for $\nu(ds_1, ds_2, A)$:

$$\alpha\, P_0(A)\,(1+\theta)\,(E_1(s_1)^{-\theta} + E_1(s_2)^{-\theta})^{-\frac{1}{\theta}-2} E_1(s_1)^{-\theta-1} E_1(s_2)^{-\theta-1} \frac{e^{-(s_1+s_2)}}{s_1 s_2} ds_1 ds_2.$$

We obtain the bound by applying Theorem 26 to the function

$$g(s_1,s_2) = E_1(s_1)^{-\theta} + E_1(s_2)^{-\theta},$$

which trivially satisfies the necessary conditions. Since g is symmetric, $K^g_{\nu,A}$ is equal to

$$2\,\alpha \int_{\mathbb{R}_+^2} \left| s_1 - (U^\pi)^{-1}\left(\int_{\mathbb{R}_+^2} \mathbb{1}_{[g(s_1,s_2),+\infty)}(g(t_1,t_2))\,\nu(t_1,t_2)\,dt_1\,dt_2 \right) \right|^2 \nu(s_1,s_2)\,ds_1\,ds_2.$$

With a change of variables $(x_1,x_2) = \left(E_1(s_1)^{-\theta}, E_1(s_2)^{-\theta}\right)$,

$$\begin{aligned}
&\int_{\mathbb{R}_+^2} \mathbb{1}_{(g(s_1,s_2),+\infty)}(g(t_1,t_2))\,\nu(t_1,t_2)\,dt_1\,dt_2 \\
=&\frac{1+\theta}{\theta^2} \int_0^{+\infty} \int_{\max(g(s_1,s_2)-t_1,0)}^{+\infty} (t_1+t_2)^{-\frac{1}{\theta}-2} dt_1\,dt_2 = \frac{1+\theta}{\theta}\, g(s_1,s_2)^{-\frac{1}{\theta}}
\end{aligned}$$

Then, with the same change of variable, the bound can be rewritten as

$$\begin{aligned}
&2\,\alpha\,\frac{1+\theta}{\theta^2} \int_0^\infty \int_0^\infty \left| E_1^{-1}\left(x_1^{-\frac{1}{\theta}}\right) - E_1^{-1}\left(\frac{1+\theta}{\theta}\,(x_1+x_2)^{-\frac{1}{\theta}}\right) \right|^2 (x_1+x_2)^{-\frac{1}{\theta}-2} dx_1\,dx_2 \\
=&2\,\alpha\,\frac{1+\theta}{\theta^2} \int_0^\infty \int_{y_1}^\infty \left| E_1^{-1}\left(y_1^{-\frac{1}{\theta}}\right) - E_1^{-1}\left(\frac{1+\theta}{\theta} y_2^{-\frac{1}{\theta}}\right) \right|^2 y_2^{-\frac{1}{\theta}-2} dy_1\,dy_2
\end{aligned}$$

We expand the binomial and treat the three terms separately.

$$\int_0^\infty \int_{y_1}^\infty E_1^{-1}\left(y_1^{-\frac{1}{\theta}}\right)^2 y_2^{-\frac{1}{\theta}-2} dy_1\,dy_2 = \frac{\theta^2}{1+\theta} \int_0^{+\infty} E_1^{-1}(x)^2 dx = \frac{\theta^2}{1+\theta}.$$

Similarly,

$$\int_0^\infty \int_{y_1}^\infty E_1^{-1}\left(\frac{1+\theta}{\theta} y_2^{-\frac{1}{\theta}}\right)^2 y_2^{-\frac{1}{\theta}-2} dy_1\,dy_2 = \frac{\theta^2}{1+\theta}$$

Thus $d_{\mathcal{W}}(\tilde{\mu}, \tilde{\mu}^{\mathrm{co}})^2 \leq 4\,c$. In order to conclude it suffices to show that

$$\lim_{\theta \to +\infty} \frac{1+\theta}{\theta^2} \int_0^\infty \int_{y_1}^\infty E_1^{-1}\left(y_1^{-\frac{1}{\theta}}\right) E_1^{-1}\left(\frac{1+\theta}{\theta} y_2^{-\frac{1}{\theta}}\right) y_2^{-\frac{1}{\theta}-2} dy_1\, dy_2 = 1.$$

Since for every $a, b \in \mathbb{R}$, $(a-b)^2 \geq 0$, the integral is smaller or equal to 1. Thus it is enough to prove it to be greater or equal to 1. We observe that since $y_1 \leq y_2$,

$$\frac{1+\theta}{\theta} y_2^{-\frac{1}{\theta}} \leq \frac{1+\theta}{\theta} y_1^{-\frac{1}{\theta}}.$$

Since E_1 is a decreasing function, so is its inverse. Thus

$$\begin{aligned}
&\frac{1+\theta}{\theta^2} \int_0^\infty \int_{y_1}^\infty E_1^{-1}\left(y_1^{-\frac{1}{\theta}}\right) E_1^{-1}\left(\frac{1+\theta}{\theta} y_2^{-\frac{1}{\theta}}\right) y_2^{-\frac{1}{\theta}-2} dy_1\, dy_2 \\
&\geq \frac{1+\theta}{\theta^2} \int_0^\infty \int_{y_1}^\infty E_1^{-1}\left(y_1^{-\frac{1}{\theta}}\right) E_1^{-1}\left(\frac{1+\theta}{\theta} y_1^{-\frac{1}{\theta}}\right) y_2^{-\frac{1}{\theta}-2} dy_1\, dy_2 \\
&= \frac{1}{\theta} \int_0^\infty E_1^{-1}\left(y_1^{-\frac{1}{\theta}}\right) E_1^{-1}\left(\frac{1+\theta}{\theta} y_1^{-\frac{1}{\theta}}\right) y_1^{-\frac{1}{\theta}-1} dy_1 \\
&= \frac{\theta}{1+\theta} \int_0^\infty E_1^{-1}\left(\frac{\theta}{1+\theta} x\right) E_1^{-1}(x) dx \\
&\geq \frac{\theta}{1+\theta} \int_0^\infty E_1^{-1}(x)^2 dx = \frac{\theta}{1+\theta},
\end{aligned}$$

which by taking the limit as $\theta \to +\infty$ is equal to 1.

3.9.13 Proof of Proposition 33

We point out that $\tilde{\boldsymbol{\xi}} \overset{\mathrm{d}}{=} \tilde{\boldsymbol{\mu}}^{\mathrm{ind}} + \tilde{\boldsymbol{\mu}}_0^{\mathrm{co}}$ and $\tilde{\boldsymbol{\xi}}^{\mathrm{co}} \overset{\mathrm{d}}{=} \tilde{\boldsymbol{\mu}}^{\mathrm{co}} + \tilde{\boldsymbol{\mu}}_0^{\mathrm{co}}$. Since by construction $\tilde{\boldsymbol{\mu}}^{\mathrm{ind}} \perp \tilde{\boldsymbol{\mu}}_0^{\mathrm{co}}$ and $\tilde{\boldsymbol{\mu}}^{\mathrm{co}} \perp \tilde{\boldsymbol{\mu}}_0^{\mathrm{co}}$, by (3.23)

$$\mathcal{W}(\tilde{\boldsymbol{\xi}}(A), \tilde{\boldsymbol{\xi}}^{\mathrm{co}}(A)) \leq \mathcal{W}(\tilde{\boldsymbol{\mu}}^{\mathrm{ind}}(A), \tilde{\boldsymbol{\mu}}^{\mathrm{co}}(A)) + \mathcal{W}(\tilde{\boldsymbol{\mu}}_0^{\mathrm{co}}(A), \tilde{\boldsymbol{\mu}}_0^{\mathrm{co}}(A)),$$

which is equal to $\mathcal{W}(\tilde{\boldsymbol{\mu}}^{\mathrm{ind}}(A), \tilde{\boldsymbol{\mu}}^{\mathrm{co}}(A))$. By taking the supremum over $A \in \mathcal{X}$ we achieve the first statement. A very similar proof can be carried on for the second, by observing that $\tilde{\boldsymbol{\xi}}^{\mathrm{ind}} \overset{\mathrm{d}}{=} \tilde{\boldsymbol{\mu}}^{\mathrm{ind}} + \tilde{\boldsymbol{\mu}}_0^{\mathrm{ind}}$.

3.9.14 Proof of Lemma 35

Let $\{A_1, \cdots A_n\}$ in $\mathcal{X}$ be disjoint sets. Then for $i = 1, \ldots, n$ the random vectors $\tilde{\boldsymbol{\mu}}_f(A_i) = \int_{A_i} f(x) \tilde{\boldsymbol{\mu}}(dx)$ are independent since f is deterministic and $\tilde{\boldsymbol{\mu}}(A_1), \cdots \tilde{\boldsymbol{\mu}}(A_n)$ are independent. This proves that $\tilde{\mu}_f$ is a CRV. The Lévy intensity ν_f may be found though the joint Laplace functional transform, $\mathbb{E}(e^{-\int g_1(x)\tilde{\mu}_1(dx) - \int g_2(x)\tilde{\mu}_2(dx)})$ for every

pair of measurable functions $g_1, g_2 : \mathbb{X} \to \mathbb{R}^+$, which characterizes the law of a CRV $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \tilde{\mu}_2)$:

$$\begin{aligned}
&\mathbb{E}\big(e^{-\int_{\mathbb{X}} g_1(x)\tilde{\mu}_{f,1}(dx) - \int_{\mathbb{X}} g_2(x)\tilde{\mu}_{f,2}(dx)}\big) = \\
&= \exp\Big\{ - \int_{\mathbb{R}_+^2 \times \mathbb{X}} [1 - e^{-(s_1 g_1(x) - s_2 g_2(x))f(x)}] \nu(ds_1, ds_2, dx) \Big\} = \\
&= \exp\Big\{ - \int_{\mathbb{R}_+^2 \times \mathbb{X}} [1 - e^{-s_1 g_1(x) - s_2 g_2(x)}] (p_f \# \nu)(ds_1, ds_2, dx) \Big\},
\end{aligned}$$

where $p_f(s_1, s_2, x) = (s_1 f(x), s_2 f(x), x)$.

3.9.15 Proof of Corollary 36

Denote $\tilde{\boldsymbol{\mu}}_t = \tilde{\boldsymbol{\mu}}_{k(t|\cdot)}$ in the notation of Lemma 35, so that $\tilde{\boldsymbol{h}}(t) \overset{\mathrm{d}}{=} \tilde{\boldsymbol{\mu}}_t(\mathbb{R})$. By definition of GM–dependence, $\tilde{\boldsymbol{\mu}}_t(dy) = k(t|y)\tilde{\boldsymbol{\mu}}^{\mathrm{ind}}(dy) + k(t|y)\tilde{\boldsymbol{\mu}}_0^{\mathrm{co}}(dy)$. If $k(t|x) = \beta \mathbb{1}_{[0,t]}(x)$ and $\tilde{\mu}$ is a gamma CRM, by Lemma 10 in Section 2.4, $k(t|x)\tilde{\mu}(dx)$ has Lévy measure

$$\pi(ds, dx) = \frac{e^{-\frac{s}{\beta}}}{s} \mathbb{1}_{(0,+\infty)}(s) \mathbb{1}_{(0,t)}(x)\, ds,$$

which corresponds to the generalized gamma CRM with parameters $b = \beta^{-1}$, $\sigma = 0$, $\alpha = t$ and $P_0 = \mathrm{Unif}([0, t])$. Thus, $\tilde{\boldsymbol{\mu}}_t$ is a special case of GM-dependent CRV with generalized gamma marginals. We conclude by Corollary 34.

Chapter 4

Posterior contraction rates of mixtures over hierarchical processes

Bayesian hierarchical models learn effectively by finding common latent features among multiple data sources. When the number of latent components can not be fixed or bounded from above, most common procedures involve depedent nonparametric priors such as the Hierarchical Dirichlet process (HDP). While the efficacy of their finite sample performance is well–established, posterior asymptotic analysis is still at an early stage and we aim at finding theoretical guarantees for the recovering of the true data generating processes. In particular, we study the posterior contraction rates for data distributions that are modeled as mixtures over the boosted hierarchical Dirichlet process, a generalization of the HDP that accomodates for a faster growth in the number of discovered latent features as the sample size increases. By extending Schwartz theory to partially exchangeable sequences we uncover the interesting behavior that posterior contraction rates depend on the relation between the numbers of observations in different groups. If these are equal or at most related in a polynomial fashion, we recover the minmax rates up to a logarithmic factor. As the relation becomes exponential, the rates may deteriorate drastically.

4.1 Introduction

The hierarchical Dirichlet process (HDP, Teh et al. (2006)) is a powerful tool for the unsupervised learning of unstructured data with a limited amount of observations. A generative process encodes the original sample space into meaningful latent features or categories. A typical example comes from information retrieval (Cowans, 2004), where the content of a collection of documents is unraveled by assigning each word of a document to one out of several topics, i.e. latent features. Topics are shared across different documents and their number is learned directly from the data, providing an effective

nonparametric version of probabilistic topic models such as the latent Dirichlet allocation (LDA, Blei et al. (2003)). These key properties have spurred numerous applications beyond information retrieval, including statistical genetics (Xing et al., 2007; Elliott et al., 2019), computer vision (Haines & Xiang, 2011), cognitive science (Griffiths et al., 2007) and robotics (Nakamura et al., 2011; Taniguchi et al., 2018), where typically the number of features can not be fixed or bounded from above. Each element of a group or subpopulation is assigned to a latent feature with an unknown probability which is learned from the data in a Bayesian way, allowing for mixed membership and borrowing of information. This is achieved through a hierarchy of Dirichlet processes (DPs) by letting the feature distribution in each subpopulation be a DP conditionally on a common parent distribution that is also a DP. Using the DP in such a hierarchical fashion brings to a slowdown in the number of new features discovered among the observations. As showed in Camerlenghi et al. (2019b), this decreases from $\log(n)$ to $\log(\log(n))$, where n is the total number of observations. To recover the $\log(n)$ growth, we consider a natural generalization of the HDP where the parent distribution is a Pitman-Yor process (PY) instead of a DP. We name such extension the boosted hierarchical Dirichlet process (bHDP) because of its ability of speeding up the growth rates of the number of features though keeping the DP at the subpopulation level.
In this chapter we find conditions for the Bayesian learning of the bHDP mixture model to recover the true generating processes of the multiple data sources. We fix the number of subpopulations and check that each distribution estimator recovers the ground truth as the number of observations diverge simultaneously at possibly different rates. Our aim is to find conditions that ensure an optimal rate of convergence (up to a logarithmic factor) according to metrics on the joint space that are built on popular distances such as the total variation and the Hellinger. This provides an asymptotic validation of the bHDP mixture model as the amount of information grows. As a by product, this brings to a new frequentist validation of the HDP mixture model, which is a special case of the bHDP. Up to our knowledge, only Nguyen (2016) dealt with this topic, though focusing on the recovery of the parent mixing distribution rather than on the true data-generating processes, and letting the number of subpopulations go to $+\infty$. In many applications the number of groups is limited by the experiment, as for example for blood type or logfiles, so that it is more reasonable to fix the number of groups and have the number of observations within each group go to infinity.
The main results are achieved through an extension of Schwartz theory to partially exchangeable sequences. Exchangeable models arise naturally from iid observations whenever one accounts for uncertainty about the parameter. Indeed by de Finetti's theorem exchangeability is equivalent to being conditionally iid. When performing inference about more than one population, the analogue of exchangeability is partial exchangeability, which allows to perform robust analysis for each group of observations by exploiting the information contained in the whole sample. Notable examples of partially exchangeable models include the HDP (Teh et al., 2006; Camerlenghi et al., 2019b), the nested Dirichlet process (Rodrguez et al., 2008; Camerlenghi et al., 2019a) and many other models built on dependent random probability measures; see Quintana et al. (2020)

for a recent review. The finite-sample performance of these Bayesian nonparametric estimators for multiple populations is well-consolidated, whereas the analysis of their asymptotic properties is still at early stage. An important frequentist validation of Bayesian inference is posterior consistency, that is the convergence of the posterior to the true distribution of the data. For exchangeable models, Schwartz theory (Schwartz, 1965) provides a general framework for dealing with consistency when the posterior distribution is not available in closed form. It relies on: (i) the existence of a sequence of tests that separates the true distribution from those outside of any neighborhood with exponentially small errors, as the number of observation grows (exponentially consistent test); (ii) sufficient prior mass on neighborhoods of the true distribution defined by the Kullback-Leibler divergence. As for (i), typically one needs to restrict to compact subsets of the parameter space. The classical frequentist theory guarantees that many notable distances, such as the Hellinger and the total variation, allow to test the true distribution against any disjoint convex set. One then proceeds as follows. First, considers a compact subset of the parameter space with most of the prior probability; second, covers the compact set with a finite number of convex balls and uses the existence of tests for convex alternatives to find a global test whose errors depend on the number of convex balls; third, shows that the number of convex balls does not grow too fast as the compact set increases, so that the global test preserves exponentially small errors. The sequence of compact subsets is usually referred to as *sieve* and the growing number of convex balls that covers it is measured in terms of *entropy*. Summarizing, one finds an exponentially consistent test by appointing a high-mass-low-entropy sieve. This is then coupled with the Kulback-Leibler support to ensure consistency and contraction rates as well. The above strategy was pioneered by Ghosal et al. (1999) to prove the consistency of the Dirichlet process mixture model (Ferguson, 1983; Lo, 1984). Following works dealt with contraction rates for densities on the real line (Ghosal & van der Vaart, 2001; Ghosal & Van Der Vaart, 2007a) and corresponding results for multivariate densities (Tokdar, 2006; Shen et al., 2013; Canale & De Blasi, 2017). As pointed out in Wu & Ghosal (2010), the extension to the multivariate setting is highly non–trivial and required the construction of a new sieve with low entropy and high mass.

Up to our knowledge, this chapter offers the first extension of Schwartz theory to partially exchangeable models. This proves crucial for finding fast contraction rates of the multivariate bHDP mixture model. Our findings may be summarized as follows: (i) to reproduce the distribution of all subpopulations simultaneously is very different from reproducing them one at a time; (ii) if all groups have the same number of observations (asymptotically), Schwartz theory for exchangeable sequences may be easily extended to partially exchangeable sequences. (iii) if the groups have a different number of observations, one has to develop a non-trivial extension of the classical theory, by requiring the prior to put sufficient mass on a *reinforced* multivariate Kullback–Leibler neighborhood, as will be made clear in Section 4.3. This is of great interest in many applied settings where the observations in each group grow at different rates. For an example concerning the blood types, it is sufficient to think that patients with type 0+ are known to be consistently more than the ones with type AB-, so that fixing to the same rate appears

as a stretch. When applied to the bHDP mixture model, the fulfillment of the reinforced Kullback–Leibler condition forces a maximum discrepancy between the smallest and the largest cardinality of the subpopulations (n and $n_\vee$ respectively). In particular when the largest increases at most polynomially with respect to the smallest, i.e. $n_\vee \lesssim n^k$ for some $k > 0$, we find an optimal rate of convergence $\sqrt{n}$ up to a logarithmic factor. If the increase is faster than polynomial, the contraction rate progressively deteriorates and fails to converge whenever $n_\vee \gtrsim e^{n^\gamma}$, where γ is a constant that depends on the dimension of the sample space and on the smoothness of the true distributions.

The chapter is organized as follows. In Section 4.2 we fix the notation and state the main result (Theorem 40) on the contraction rates for the bHDP mixture model. The following sections develop the analytical tools for achieving Theorem 40. In Section 4.3 we extend Schwartz theory to partially exchangeable models, with a particular focus on the case of different cardinalities between groups. This is applied to the bHDP mixture model in Section 4.4, whereas future applications are discussed in Section 4.5. All proofs are deferred to Section 4.6.

4.2 Preliminaries and main result

In this section we define the boosted hierarchical Dirichlet process mixture model and fix the notation for partially exchangeable models and their posterior convergence rates. This will enable to state the main result of the chapter (Theorem 40), whose proof is built in Section 4.3 and Section 4.4.
Let $\alpha, \theta > 0$ and $F \in \mathcal{P}$, where $\mathcal{P} = \mathcal{P}_{\mathbb{X}}$ indicates the set of probability distributions on a Polish space $\mathbb{X}$. We indicate with $\mathrm{DP}(\theta, F)$ the Dirichlet process (Ferguson, 1983) and $\mathrm{PY}(\alpha, \theta, F)$ the Pitman–Yor process or two-parameter Poisson–Dirichlet process (Pitman & Yor, 1997). The boosted hierarchical Dirichlet process $(F_i)_{i=1}^m \sim \mathrm{bHDP}(\theta, \alpha^*, \theta^*, F^*)$ is defined as

$$\begin{aligned} &(F_i)_{i=1}^m | \tilde{F} \overset{\text{iid}}{\sim} \mathrm{DP}(\theta, \tilde{F}); \\ &\tilde{F} \sim \mathrm{PY}(\alpha^*, \theta^*, F^*), \end{aligned} \tag{4.1}$$

where $\theta, \theta^* > 0$, $\alpha^* \geq 0$ and $F^* \in \mathbb{P}$. By letting $\alpha^* = 0$ we recover the hierarchical Dirichlet process (HDP) as a special case. We consider the bHDP as nonparametric prior for the following Gaussian mixture model. Let $m, n_1, \dots, n_m \in \mathbb{N} \setminus \{0\}$. A sequence $\boldsymbol{X} = (X_{i,j})_{(i,j)}$ with $i = 1, \dots, m$ and $j = 1, \dots, n_i$ on a Polish space $\mathbb{X}$ satisfies $\boldsymbol{X} \sim \mathrm{bHDPM}(G, \theta, \alpha^*, \theta^*, F^*)$ if

$$\begin{aligned} &(X_{i,1}, \dots, X_{i,n_i})_{i=1}^m | (F_i)_{i=1}^m, (\Sigma_i)_{i=1}^m \sim P_{F_1,\Sigma_1}^{(n_1)} \times \cdots \times P_{F_m,\Sigma_m}^{(n_m)}, \\ &((F_i)_{i=1}^m, (\Sigma_i)_{i=1}^m) \sim \Pi := \mathrm{bHDP}(\theta, \alpha^*, \theta^*, F^*) \times G^{(m)}, \end{aligned} \tag{4.2}$$

where P_{F_i,Σ_i} probability distribution of the multivariate Gaussian mixture, whose density function $p_{F,\Sigma}$ is defined as follows. Let ϕ_Σ be the density of the d–dimensional Gaussian

distribution centered in the origin and with covariance matrix Σ, where Σ is a positive–definite matrix of dimension $d \times d$ and let $F \in \mathcal{P}_{\mathbb{R}^d}$. Then,

$$p_{F,\Sigma}(x) = \int_{\mathbb{R}^d} \phi_\Sigma(x-y)\, dF(y) = (\phi_\Sigma * F)(x), \tag{4.3}$$

where $x = (x_1, \dots, x_d)$ and $y = (y_1, \dots, y_d)$. We call F the mixing distribution.
The bHDPM model induces partially exchangeable observations. We recall that $\boldsymbol{X}$ is partially exchangeable if there exists an m–dimensional vector of random probabilities $(\tilde{P}_1, \dots, \tilde{P}_m)$ such that

$$\begin{aligned} &(X_{i,1}, \dots, X_{i,n_i})_{i=1}^m | (\tilde{P}_1, \dots, \tilde{P}_m) \sim \tilde{P}_1^{(n_1)} \times \cdots \times \tilde{P}_m^{(n_m)}, \\ &(\tilde{P}_1, \dots, \tilde{P}_m) \sim \Pi, \end{aligned} \tag{4.4}$$

where $\tilde{P}^{(n)} = \prod_{i=1}^n \tilde{P}$ indicates the product probability and Π is a probability on the product space of probabilities $\mathcal{P}^m = \mathcal{P}_{\mathbb{X}}^m$. In particular, each group of observations is exchangeable and satisfies $(X_{i,1}, \dots, X_{i,n_i}) | \tilde{P} \sim \tilde{P}; \tilde{P} \sim \Pi_i$, where Π_i is the i-th marginal distribution, i.e. $\Pi_i(A) = \Pi(\mathcal{P}^{i-1} \times A \times \mathcal{P}^{m-i})$ for any measurable $A \subset \mathcal{P}$, with the convention $\mathcal{P}^0 = \emptyset$.
We focus on partially exchangeable models (4.4) such that Π has support over absolutely continuous distributions with respect to some measure λ on $\mathbb{X}$, so that $\tilde{P}_i$ has density $\tilde{p}_i$ almost surely, for every $i = 1, \dots, m$. We indicate by $\Pi(\cdot|\boldsymbol{X})$ a version of the posterior distribution according to model (4.4). We then assume that $(X_{i,1}, \dots, X_{i,n_i})_{i=1}^m \sim P_{0,1}^{(n_1)} \times \cdots \times P_{0,m}^{(n_m)}$ are m independent groups of observations, where $P_{0,i}$ is absolutely continuous with respect to λ with density $p_{0,i}$, for $i = 1, \dots, m$. We analyze the properties of $\Pi(\cdot|\boldsymbol{X})$ as an estimator of $(P_{0,i})_{i=1}^m$, as the number of groups is fixed and the number of observations goes to $+\infty$ in each group, with possibly different orders of magnitude. To this end, we define $n = \min(n_1, \dots, n_m)$ and take the limit as $n \to +\infty$.
We recall that every topology on $\mathcal{P}$ is inherited by the product space $\mathcal{P}^m$ through the product topology. We say that a metric $d^{(m)}$ on $\mathcal{P}^m$ is a product metric whenever it metrizes the product topology. If d is a metric on $\mathcal{P}$ defined on densities, one of the most notable classes of product metrics are the ℓ_p–metrics, for $1 \leq p < +\infty$:

$$d_p((p_i)_{i=1}^m, (q_i)_{i=1}^m) = \left(\sum_{i=1}^m d(p_i, q_i)^p \right)^{\frac{1}{p}}. \tag{4.5}$$

Definition 12. A sequence ϵ_n is a posterior contraction rate at $(P_{0,i})_{i=1}^m$ with respect to d_p in (4.5) if

$$\Pi(d_p((p_i)_{i=1}^m, (p_{0,i})_{i=1}^m) \geq M_n \epsilon_n | \boldsymbol{X}) \to 0$$

in $(\prod_{i=1}^m P_{0,i}^{(\infty)})$–probability, for every $M_n \to +\infty$, as $n \to +\infty$.

We find contraction rates towards supersmooth density function. Denote by $[-z, z]^d = \prod_{i=1}^d [-z, z]$ be the d-dimensional cube of side $[-z, z]$.

Definition 13. A density function p_0 on $\mathbb{R}^d$ is said to be *supersmooth* if there exist (F_0, Σ_0) such that $p_0 = p_{F_0,\Sigma_0}$ and $1 - F_0([-z,z]^d) \lesssim e^{-c_0 z^{r_0}}$ for every $z > 0$, with $c_0 > 0$ and $r_0 \geq 2$. We refer to r_0 as the smoothness parameter.

We make the following standard assumption on the parameters F^* and G of the bHDPM model in (4.2): there exist positive constants a_k, C_k, b_k, for $k = 1, 2, 3$, such that

$$1 - F^*([-z,z]^d) \leq b_1 e^{-C_1 z^{a_1}}, \tag{4.6}$$
$$G(\Sigma : \mathrm{eig}_d(\Sigma^{-1}) \geq s) \leq b_2 e^{-C_2 s^{a_2}}, \qquad G(\Sigma : \mathrm{eig}_1(\Sigma^{-1}) \leq s) \leq b_3 s^{a_3}.$$

Theorem 40 (Main result)**.** *Let $\Pi(\cdot|\boldsymbol{X})$ indicate the posterior distribution according to the model $\boldsymbol{X} \sim \mathrm{bHDPM}(\mathrm{G}, \theta, \alpha^*, \theta^*, \mathrm{F}^*)$ in (4.2) such that conditions (4.6) hold. Let $p_{0,i} = p_{F_{0,i},\Sigma_{0,i}}$ be a supersmooth density with $\Sigma_{0,i}$ in the support of G and smoothness $r_{0,i}$, for $i = 1, \ldots, m$ and denote by $r_0 = \min(r_{0,1}, \ldots, r_{0,m})$. Assume that there exists $0 < \delta < (d + d/r_0 + 2)^{-1}$ such that $n_\vee \lesssim e^{n^\delta}$ for n large enough. Then $n^{-1/2} \log(n_\vee)^{(d+d/r_0+2)/2}$ is a posterior contraction rate at $(p_{0,i})_{i=1}^m$ with respect to d_p for every $p \geq 1$, where d is the Hellinger or the total variation distance.*

Theorem 40 focuses on a partially exchangeable model whose marginal exchangeable sequences have optimal rates of convergence up to a logarithmic factor, as shown in the proof of Theorem 40. We observe that the marginal exchangeable sequences are not mixtures over a Dirichlet process, since the prior in this context is rather

$$\Pi_i(\cdot) = \mathbb{E}(\mathrm{DP}_{\theta,\tilde{F}}(\cdot)) = \int_{\mathcal{P}_{\mathbb{X}}} \mathrm{DP}_{\theta,\tilde{F}}(\cdot)\, d\mathrm{PY}_{\alpha^*,\theta^*,F^*}(\tilde{F}),$$

where $\mathrm{DP}_{\theta,\tilde{F}} = \mathrm{DP}(\theta, \tilde{F})$ and $\mathrm{PY}_{\theta^*,\alpha^*,F^*} = \mathrm{PY}(\theta^*, \alpha^*, F^*)$. As a by product, we have thus found optimal convergence rates for a new set of exchangeable sequences with non-parametric priors. Moreover, Theorem 40 provides contraction rates for the partially exchangeable sequence given by the bHDPM process defined in (4.2) as the number of observations in each group grow at different speeds. In particular, we observe that if the largest group grows at a polynomial speed with respect to the smallest group, i.e. $n_\vee \lesssim n^k$ for some $k > 0$, the contraction rate is optimal up to a logarithmic factor with respect to the cardinality of the smallest group, i.e. $n^{-1/2} \log(n)^{(d+d/r_0+2)/2}$ is a contraction rate. However, when the growth becomes exponentially fast the contraction rate deteriorates progressively, becoming non–informative whenever $n_\vee \gtrsim e^{n^{1/(d+d/r_0+2)}}$. It is of interest to underline the doubly positive effect of the smoothness parameter: as r_0 increases, it makes both the contraction rate faster and the range of growth rates of $n_\vee$ wider. As an illustrating, consider two groups of observations having cardinality n_1, n_2 respectively. Assume that $(X_{i,1}, \ldots, X_{i,n_i})_{i=1}^2 \sim \mathrm{Norm}(a_1, \Sigma_1)^{(n_1)} \times \mathrm{Norm}(a_2, \Sigma_2)^{(n_2)}$, where $\mathrm{Norm}(a, \Sigma)$ is a bivariate Gaussian distribution with mean vector a and covariance matrix Σ, so that $d = 2$ and $r_0 = +\infty$. Then as n_1 and n_2 diverge, assuming without loss of generality that $n_1 \lesssim n_2$, the posterior distribution of the densities corresponding to the bHDPM model contracts towards the vector of true Gaussian distributions whenever $n_2 \lesssim e^{{n_1}^{1/4}}$, with a rate of contraction equal to $\sqrt{\log(n_2)^4/n_1}$. On the contrary,

when $n_2 \gtrsim e^{{n_1}^{1/4}}$, the convergence to the truth is not ensured.
We underline that in general our results hold when the true distributions are supersmooth, though these could be adequately extended to smooth densities by placing some Hölder conditions on the derivatives. We decided to focus on supersmooth functions because all the crucial elements in the asymptotic analysis of partially exchangeable models are already present in this simplified context. Hopefully, this allows to untangle the technicalities due to the presence of multiple populations from the ones due to the smoothness assumptions on the true densities.

We conclude this section with some further notation. If $P \in \mathcal{P}$ and $\phi : \mathbb{X} \to \mathbb{R}$ is a measurable function, $P(\phi)$ indicates the expected value of ϕ with respect to the probability $P \in \mathcal{P}$. We use the symbol $\overset{\text{ind}}{\sim}$ for independent random variables and $\overset{\text{iid}}{\sim}$ for independent and identically distributed ones. Let $(a_n)_{n\in\mathbb{N}}, (b_n)_{n\in\mathbb{N}}$ two sequences on $\mathbb{R}$. We write $a_n \asymp b_n$ if they are of the same order as a function of n, i.e. $a_n b_n^{-1} \to K$ as $n \to +\infty$, where K is a constant different from zero. If $K = 0$, we use the notation $a_n \ll b_n$. Moreover, we write $a_n \lesssim b_n$ if a_n is smaller than b_n up to an irrelevant constant. Similarly for $a_n \gtrsim b_n$. The Lebesgue measure on $A \subset \mathbb{R}^n$ is denoted by $\mathscr{L}_n(A)$. The negative part of $\log(\cdot)$ is denoted by $\log_-(\cdot) = \max(-\log(\cdot), 0)$. The ascending factorial of $\beta \in \mathbb{R}$ is $\beta^{[n]} = \beta(\beta+1)\dots(\beta+n-1)$. Given a vector $(a_1, \dots, a_m)$, we write $a = \min(a_1, \dots, a_m)$, $a_\vee = \max(a_1, \dots, a_m)$ and $a_+ = a_1 + \dots + a_m$. Given two sets $A \subset B$, we indicate with $A^c = B \setminus A$ the complement of A in B. The cardinality of A is denoted by $|A|$. The Kullback–Leibler divergence between two densities p_1, p_2 is $\mathrm{KL}(p_1; p_2) = P_1(\log(p_1/p_2))$; the Hellinger distance is $d_H(p_1, p_2) = \int(\sqrt{p_1} - \sqrt{p_2})^2 \, d\lambda$; the total variation distance is $\mathrm{TV}(p_1, p_2) = 2^{-1} \int |p_1 - p_2| \, d\lambda$. The L_p–norm of a density q is $\|q\|_p = (\int |q|^p \, d\lambda)^{p^{-1}}$, so that $\mathrm{TV}(p_1, p_2) = 2^{-1}\|p_1 - p_2\|_1$. Similarly, the ℓ_p–norm in $\mathbb{R}^n$ is $\|(q_i)_{i=1}^n\|_p = (\sum_{i=1}^n |q_i|^p)^{p^{-1}}$. The convolution of two measurable functions f_1, f_2 is denoted by $(f_1 * f_2)(x) = \int f_1(x-y) f_2(y) \, dy$. For any $\epsilon > 0$, a subset S of a metric space (T, d) is an ϵ-net if for every $t \in T$ there exists $s \in S$ such that $d(s, t) < \epsilon$. We call the ϵ-covering of T the minimal cardinality of ϵ-nets S and denote such number by $\mathscr{N}(\epsilon, T, d)$. We refer to a test as any measurable function $\phi : \mathbb{X}^k \to [0, 1]$ for some $k \in \mathbb{N}$.

4.3 Contraction rates for partially exchangeable sequences

In this section we prove a general Schwartz theorem for partially exchangeable sequences that we apply to the boosted hierarchical Dirichlet process in Section 4.4. In Theorem 41 we find general conditions for the convergence rate of $(p_i)_{i=1}^n | \boldsymbol{X}$ to be deduced from the marginal ones of $p_i | X_{i,1}, \dots X_{i,n_i}$, for $i = 1, \dots, m$, which is particularly delicate whenever the cardinalities of the groups do not all grow at the same rate. We observe that the connection between marginal convergence rates and joint ones is not trivial, since in general posterior consistency for the marginal exchangeable sequences does not imply the one for the partially exchangeable ones, as it is evident from Example 6.

Example 6. We focus on consistency with respect to the weak topology. Let $m = 2$ and

let Π be a prior whose support is $\{p_1 \text{ s.t. } \mathrm{KL}(p_{0,1}, p_1) \geq \tilde{\epsilon}\} \times \mathcal{P} \cup \mathcal{P} \times \{p_2 \text{ s.t. } \mathrm{KL}(p_{0,2}, p_2) \geq \tilde{\epsilon}\}$ for some $\tilde{\epsilon} > 0$ such that the following holds for every $\epsilon > 0$:

$$\int_{\{\mathrm{KL}(p_{0,1},p_1)<\epsilon\}\times\mathcal{P}} \Pi(dp_1, dp_2) > 0, \qquad \int_{\mathcal{P}\times\{\mathrm{KL}(p_{0,2},p_2)<\epsilon\}} \Pi(dp_1, dp_2) > 0. \tag{4.7}$$

Let $\mathcal{U}_0 = \{\mathrm{KL}(p_{0,1}, p_1) < \tilde{\epsilon}\} \times \{\mathrm{KL}(p_{0,1}, p_1) < \tilde{\epsilon}\}$. Since $\Pi(\mathcal{U}_0) = 0$, the posterior according to the partially exchangeable model (4.4) satisfies $\Pi(\mathcal{U}_0|\boldsymbol{X}) = 0$, $\mathcal{L}(\boldsymbol{X})$–almost surely, for every $n_1, n_2 \in \mathbb{N} \setminus \{0\}$. Since $\mathcal{L}(\boldsymbol{X})$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{n_1+n_2}$, $\Pi(\mathcal{U}_0|\boldsymbol{X}) = 0$ $\left(P_{0,1}^{(n_1)} \times P_{0,2}^{(n_2)}\right)$–almost surely and thus also in probability. As $\mathcal{U}_0$ is a neighborhood of $(P_{0,1}, P_{0,2})$ according to the weak topology, Π is not consistent at $(P_{0,1}, P_{0,2})$. However, (4.7) guarantees that the marginal random measures Π_1 and Π_2 satisfy the KL–property of Schwartz theorem for exchangeable sequences, ensuring marginal consistency (see e.g. Ghosal & van der Vaart (2017, Example 6.20)).

We assume that the base metric of the ℓ_p–metric d_p defined in (4.5) satisfies the basic testing assumption: given $p_0 \in \mathcal{P}$, for every $n \in \mathbb{N}, \epsilon > 0$ and $p_1 \in \mathcal{P}$ such that $d(p_0, p_1) > \epsilon$, there exists a test $\tilde{\phi}_n : \mathbb{X}^n \to [0, 1]$ and some universal constants $\xi, C > 0$ such that

$$P_0^n(\tilde{\phi}_n) \leq e^{-Cn\epsilon^2}; \qquad \sup_{d(p,p_1)<\xi\epsilon} P^n(1 - \tilde{\phi}_n) \leq e^{-Cn\epsilon^2}. \tag{4.8}$$

This standard requirement holds for the Hellinger distance (Le Cam, 1986). More generally, it holds for any metric $d \leq d_H$ that generates convex balls (cfr. Proposition D.8 in Ghosal & van der Vaart (2017)), including the total variation distance. We define the reinforced Kullback–Leibler variation neighborhood

$$\mathcal{V}_{0,\epsilon,n} = \left\{ \mathrm{KL}(p_{0,i}; p_i), V(p_{0,i}; p_i) \lesssim \frac{n}{n_i}\epsilon^2 \text{ for } i = 1, \dots m \right\}, \tag{4.9}$$

where $V(p; q) = P|\log(p/q) - \mathrm{KL}(p; q)|^2$ is the Kullback–Leibler variation. We observe that (4.9) differs from the standard definition of Kullback–Leibler variation neighborhood $\mathcal{V}_{0,\epsilon} = \{\mathrm{KL}(p_{0,i}; p_i), V(p_{0,i}; p_i) \lesssim \epsilon^2 \text{ for } i = 1, \dots m\} \supset \mathcal{V}_{0,\epsilon,n}$, introducing an explicit dependence on the cardinality of the samples (not only through the convergence rate), which has the effect of shrinking each component $\{\mathrm{KL}(p_{0,i}; p_i), V(p_{0,i}; p_i) \lesssim \epsilon^2\}$ of the neighborhood proportionally to the ratio between $n = \min(n_1, \dots, n_m)$ and n_i. We added a subscript n in the notation $\mathcal{V}_{0,\epsilon,n}$, though technically it also depends on the whole vector $(n_1, \dots, n_m)$. We observe that when $n_i \asymp n_{i'}$ for every $i \neq i'$, $\mathcal{V}_{0,\epsilon,n}$ and $\mathcal{V}_{0,\epsilon}$ coincide.

Theorem 41. *Given a distance d that satisfies the basic testing assumption (4.8), suppose that there exist $\mathcal{P}_n \subset \mathcal{P}$ and a constant $C > 0$, such that for $\bar{\epsilon}_n \leq \epsilon_n$ sequences of real numbers such that $n\bar{\epsilon}_n^2 \to +\infty$, the following hold for sufficiently large n:*

1. $\Pi(\mathcal{V}_{0,\bar{\epsilon}_n,n}) \geq e^{-Cn\bar{\epsilon}_n^2}$;

2. $\log(\mathscr{N}(\xi\epsilon_n, \mathcal{P}_n, d)) \leq n\epsilon_n^2$;

3. $\Pi_i(\mathcal{P}_n^c) \leq e^{-(C+2m+1)n\bar{\epsilon}_n^2}$ *for* $i = 1, \dots, m$.

Then ϵ_n *is a posterior rate of contraction at* $(p_{0,i})_{i=1}^m$ *with respect to* d_p, *for every* $p \geq 1$.

The proof of Theorem 41 may be found in Section 4.6 but we give some intuition on the role of the conditions and their relation with the exchangeable case. The basic testing assumption (4.8) and conditions 2 and 3 of Theorem 41 are standard assumptions for building marginal frequentist tests $\phi_{n_i}^i : \mathbb{X}^{n_i} \to \{0,1\}$ that separate the true distribution $P_{0,i}$ from the complementary of any neighborhood with exponentially small errors with respect to the number of observations. In the statement of Theorem 41 we considered the same sieve $(\mathcal{P}_n)_{n\geq 1}$ for every marginal distribution Π_i for simplicity, since in most common frameworks, including the boosted hierarchical Dirichlet process in (4.1), $\Pi_i = \Pi_{i'}$ for every i, i'. However, we point out that the result may be generalized to account for different subsets $\mathcal{P}_{n,i} \subset \mathcal{P}$, as showed in Section 4.6, which is of particular interest when the marginal exchangeable models with respect to Π_i require different sieves.
Condition 1 on the reinforced Kullback–Leibler variation neighborhood is needed because one can not directly build a frequentist test $\phi : \mathbb{X}^{n_+} \to \{0,1\}$ that separates the true distributions $(P_{0,i})_{i=1}^m$ with exponentially bounded errors with respect to the total number of observations n_+, unless $n_i \asymp n_{i'}$ for every i, i'. However, in Lemma 42 we manage to build an exponentially bounded test with respect to n. Given tests $\{\phi^i\}_{i=1}^m$, we define

$$\phi := \sum_{k=1}^{m} (-1)^{i-1} \sum_{I \in \mathcal{I}_{m,k}} \prod_{i \in I} \phi_{n_i}^i,$$

where $\mathcal{I}_{m,k} = \{I \subset \{1, \dots, m\} : |I| = k\}$.

Lemma 42. *Let* $\mathbb{X}_1, \dots \mathbb{X}_m$ *be Polish spaces and let* $P_{0,i} \in \mathcal{P}_{\mathbb{X}_i}$ *for* $i = 1, \dots, m$. *Given a neighborhood* $\mathcal{U}_i$ *of* $P_{0,i}$ *and a measurable subset* $\mathcal{P}_i \subset \mathcal{P}_{\mathbb{X}_i}$, *assume that there exists a constant* a_i *and a test* $\phi_i : \mathbb{X}_i \to [0,1]$ *such that*

$$P_{0,i}(\phi_i) < e^{-a_i}, \qquad \sup_{p \in \mathcal{P}_i \cap \mathcal{U}_{0,i}^c} P(1-\phi_i) < e^{-a_i},$$

for $i = 1, \dots, m$. *Then* $\phi : \prod_{i=1}^m \mathbb{X}_i \to [0,1]$ *is a test and satisfies*

1. $\big(\prod_{i=1}^m P_{0,i}\big)(\phi) < me^{-a}$;

2. $\sup_{(p_i)_{i=1}^m \in \mathcal{P}_{(m)} \cap \mathcal{U}_0^c} \big(\prod_{i=1}^m P_i\big)(1-\phi) < e^{-a}$;

where $\mathcal{P}_{(m)} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_m$, $\mathcal{U}_0 = \mathcal{U}_{0,1} \times \cdots \times \mathcal{U}_{0,m}$ *and* $a = \min(a_1, \dots, a_m)$.

We apply Lemma 42 to marginal frequentist test $\phi_{n_i}^i : \mathbb{X}^{n_i} \to \{0,1\}$ that are exponentially bounded with respect to the number of observations n_i. The fact that ϕ is exponentially bounded with respect to n instead of n_+ brings to the need of the reinforced Kullback–Leibler condition, which coincides with the standard one whenever

$n_i \asymp n_{i'}$ for every i, i'. Indeed, one can show that in such case Lemma 43 holds, whereas with the standard Kullback–Leibler variation one would only retain a lower bound in n_+.

Lemma 43. *Consider model (4.4) when Π is supported on dominated distributions. Then for any $\epsilon, D > 0$, for n sufficiently large,*

$$\int \prod_{i=1}^{m} \prod_{j=1}^{n_i} \frac{p_i}{p_{0,i}}(X_{i,j})\, d\Pi(p_1, \dots, p_m) \geq \Pi(\mathcal{V}_{0,\epsilon,n}) e^{-m(D+1)\epsilon^2 n}, \tag{4.10}$$

with $\left(\prod_{i=1}^{m} P_{0,i}^{(\infty)}\right)$–probability at least $1 - (mD^2\epsilon^2 n)^{-1}$.

4.4 Boosted hierarchical Dirichlet process

In this section we apply Theorem 41 to find the contraction rates of the multivariate Gaussian bHDPM process defined in (4.1) towards the true vector of densities $(p_{0,i})_{i=1}^m$. The conditions of Theorem 41 are divided in two separate blocks: (i) the Kullback–Leibler variation neighborhood must have sufficient mass (condition 1); (ii) one must find an appropriate sieve for the marginal exchangeable model (condition 2 and 3). Lemma 44, Proposition 45 and Proposition 46 deal with (i), Proposition 47 deals with (ii).
In order to prove that condition 1 of Theorem 41 holds, we first show that the prior puts sufficient mass on a family of neighborhood B_ϵ that contains the Hellinger ball (Lemma 44 and Proposition 45) and then show that this is sufficient to ensure that the Kullback–Leibler variation neighborhood has sufficient mass with respect to $\bar{\epsilon_n}^2 = n^{-1}\log(n_\vee)^\gamma$ (Proposition 46), for some $\gamma > 1$, which is optimal up to a logarithmic factor.

Lemma 44. *Let $p_{0,i} = p_{F_{0,i},\Sigma_{0,i}}$ be a supersmooth density on $\mathbb{R}^d$ of smoothness $r_{0,i}$, as defined in (13), and let $p_i = p_{F_i,\Sigma_i}$ be Gaussian mixture densities as defined in (4.3), for $i = 1, \dots, m$. Then for $i = 1, \dots, m$ there exist measurable subsets $\{U_{i,j}\}_{j=1}^{N_i} \subseteq \mathbb{R}^d$ and weights $\{\omega_{i,j}\}_{j=1}^{N_i} \in \mathbb{S}_{N_i-1}$ such that the following hold for sufficiently small $\epsilon > 0$:*

1. *$N_i \lesssim \log_-(\epsilon)^{d+d/r_0}$;*
2. *$U_{i,j} \cap U_{i',j'} \neq \emptyset$ if and only if $i \neq i'$ and $U_{i,j} = U_{i',j'}$;*
3. *$\mathrm{diam}(\mathrm{U_{i,j}}) \geq \epsilon^2$ for $i = 1, \dots, m$ and $j = 1, \dots, N_i$;*
4. *$\{d_H(p_{0,i}, p_i) \lesssim \epsilon \text{ for } i = 1, \dots, m\} \supseteq B_\epsilon$,*

where $B_\epsilon = \{\sum_{j=1}^{N_i} |F_i(U_{i,j}) - \omega_{i,j}| \leq \epsilon^2, \|\Sigma_i - \Sigma_{0,i}\| \leq \epsilon \text{ for } i = 1, \dots, m\}$ and $r_0 = \min(r_{0,1}, \dots, r_{0,m})$.

Proposition 45. *Let* $\Pi = \text{bHDP}(\theta, \alpha^*, \theta^*, \text{F}^*) \times \text{G}^{(\text{m})}$ *as in* (4.2). *Let* $p_{0,i} = p_{F_{0,i},\Sigma_{0,i}}$ *be a supersmooth density on* $\mathbb{R}^d$ *of smoothness* $r_{0,i}$ *such that* $\Sigma_{0,i}$ *belongs to the support of* G, *for* $i = 1, \dots, m$. *Let* B_ϵ *be the set defined in Lemma 44. Then there exists a constant* $C > 0$ *such that for sufficiently small* $\epsilon > 0$, $\Pi(B_\epsilon) \geq e^{-C \log_-(\epsilon)^{d+d/r_0+1}}$, *where* $r_0 = \min(r_{0,1}, \dots, r_{0,m})$.

We state Proposition 46 for the bHDP. However, the result holds true for a generic prior $((F_i)_{i=1}^m, (\Sigma_i)_{i=1}^m) \sim \Pi_m \times G^{(m)}$. One may thus use Proposition 46 to find the convergence rates of other dependent random probabilities, going beyond the bHDP and other hierarchical processes.

Proposition 46. *Let* $\Pi = \text{bHDP}(\theta, \alpha^*, \theta^*, \text{F}^*) \times \text{G}^{(\text{m})}$ *as in* (4.2) *and let* $p_{0,i} = p_{F_{0,i},\Sigma_{0,i}}$ *be a supersmooth density on* $\mathbb{R}^d$, *for* $i = 1, \dots, m$. *Assume that there exists* $\gamma > 2$ *and* $C > 0$ *such that, for sufficiently small* $\epsilon > 0$, $\Pi(B_\epsilon) \geq e^{-C \log_-(\epsilon)^\gamma}$, *where* B_ϵ *is the set defined in Lemma 44. If* $\bar{\epsilon_n}^2 = n^{-1} \log(n_\vee)^\gamma \to 0$ *as* $n \to +\infty$, *then there exists* $C' > 0$ *such that, for sufficiently large* n, $\Pi(\mathcal{V}_{0,\bar{\epsilon}_n,n}) \geq e^{-C' n \bar{\epsilon}_n^2}$, *where* $\mathcal{V}_{0,\epsilon,n}$ *is the Kullback–Leibler variation neighborhood defined in* (4.9).

The next proposition provides a sieve that satisfies the desired conditions, in the same spirit of Shen et al. (2013). We point out that this holds for hierarchical models with conditionally Dirichlet marginals and mean measure $F^*(A) = \mathbb{E}(\tilde{F}(A))$ in general, regardless of the prior on $\tilde{F}$. First of all we define some relevant quantities.

$$\mathcal{F}_{N,a} = \left\{ \sum_{j=1}^{+\infty} \omega_j \delta_{z_j} \middle| \sum_{j=N+1}^{+\infty} \omega_j < \epsilon_n^2, z_1, \dots, z_N \in [-a, a]^d \right\};$$

$$\mathcal{S}_{\sigma,M} = \{\Sigma : \sigma^2 \leq \text{eig}_1(\Sigma) \leq \text{eig}_d(\Sigma) < \sigma^2(1 + \epsilon_n^2)^M\};$$

$$\bar{\epsilon_n}^2 = n^{-1} \log(n_\vee)^\gamma; \qquad \epsilon_n^2 = \frac{K^2 \bar{\epsilon}_n^2 \log(n)}{\log(n \bar{\epsilon}_n^2)}; \qquad N_n = \frac{K n \bar{\epsilon}_n^2}{\log(n \bar{\epsilon}_n^2)};$$

$$a_n = (n\epsilon_n^2)^{\frac{1}{\alpha_1}}; \qquad \sigma_n^{-1} = (n\epsilon_n^2)^{\frac{1}{2\alpha_2}}; \qquad M_n = n.$$

Proposition 47. *Let* $((F_i)_{i=1}^m, (\Sigma_i)_{i=1}^m) \sim \Pi = \text{bHDP}(\theta, \alpha^*, \theta^*) \times \text{G}^{(\text{m})}$ *as in* (4.2) *such that conditions* (4.6) *hold. Define* $\mathcal{P}_n = \{p_{F_n,\Sigma_n} : F \in \mathcal{F}_{N_n,a_n}, \Sigma \in \mathcal{S}_{\sigma_n,M_n}\}$ *for* $K > 0$. *If* $\bar{\epsilon_n}^2 = n^{-1} \log(n_\vee)^\gamma \to 0$ *as* $n \to +\infty$, *then* $\log(\mathcal{N}(\epsilon_n, \mathcal{P}_n, d)) \leq n\epsilon_n^2$ *and for every* $C > 0$ *there exists* $K > 0$ *such that* $\Pi_i(\mathcal{P}_n^c) \leq e^{-Cn\bar{\epsilon}_n^2}$ *for every* $i = 1, \dots, m$.

Putting together Lemma 44, Proposition 45, Proposition 46 and Proposition 47, we obtain the proof of Theorem 40.

Proof of Theorem 40 Let $\bar{\epsilon}_n^2 = n^{-1} \log(n_\vee)^{d+d/r_0+1}$ and $\epsilon_n^2 = n^{-1} \log(n_\vee)^{d+d/r_0+2}$. We point out that if $n_\vee \leq e^{n^\delta}$, $\bar{\epsilon}_n, \epsilon_n \to 0$ as $n \to +\infty$. By relying on Proposition 45, Condition 1 on the reinforced Kullback–Leibler variation holds by Proposition 46. Consider now $\mathcal{P}_n$ as in Proposition 47 and denote with $\tilde{\epsilon}_n$ the value of ϵ_n therein. For n large enough, $\tilde{\epsilon}_n \lesssim \epsilon_n$, so that condition 2 holds by Proposition 47. Finally, $\Pi_i(\mathcal{P}_n^c) \leq e^{-Cn\bar{\epsilon}_n}$ for every $i = 1, \dots, m$ and $C > 0$. In particular, condition 3 holds as well.

4.5 Future developments

In this chapter we have laid the groundwork for the analysis of the frequentist properties of models involving dependent random probability measures, such as the boosted hierarchical Dirichlet process. In principle the same techniques can be used for the entire class of hierarchical Pitman–Yor mixture models (Camerlenghi et al., 2019b), where both the child and the parent distribution are PYs instead than DPs. However, in order to treat this class we first need an exhaustive asymptotic theory for the exchangeable Pitman–Yor mixture model, which is currently missing in the multivariate scenario. The sieve proposed in Shen et al. (2013) for the multivariate Dirichlet process mixtures is inherently dependent on fast decreasing weights, leaving the Pitman–Yor case currently unresolved. Still, we may use Theorem 41 to derive contraction rates in this class of models by adding more restrictive assumptions, such as real-valued observations (Scricciolo, 2014) or multivariate distributions with compact support.

4.6 Proofs

4.6.1 Proof of Theorem 41

We prove a more general statement of Theorem 41, which accounts for potentially different sieves for each marginal distribution.

Theorem 48. *Given a distance d that satisfies the basic testing assumption (4.8), suppose that there exist $\mathcal{P}_i = \mathcal{P}_{n,i} \subset \mathcal{P}$ and a constant $C > 0$, such that for $\bar{\epsilon}_n \leq \epsilon_n$ sequences of real numbers such that $n\bar{\epsilon}_n^2 \to +\infty$, the following hold for sufficiently large n:*

1. $\Pi(\mathcal{V}_{0,\bar{\epsilon}_n,n}) \geq e^{-Cn\bar{\epsilon}_n^2}$;
2. $\log(\mathscr{N}(\xi\epsilon_n, \mathcal{P}_i, d)) \leq n_i\epsilon_n^2$ *for* $i = 1, \dots, m$;
3. $\Pi_i(\mathcal{P}_i^c) \leq e^{-(C+2m+1)n\bar{\epsilon}_n^2}$ *for* $i = 1, \dots, m$.

Then ϵ_n is a posterior rate of contraction at $(p_{0,i})_{i=1}^m$ with respect to d_p, for every $p \geq 1$.

Proof. Let $B_n = \{(p_i)_{i=1}^m \in \mathcal{P}^m : d_p((p_i)_{i=1}^m, (p_{0,i})_{i=1}^m) > M\epsilon_n\}$. Since convergence in expected value implies convergence in probability, we shall prove that $\mathbb{E}(\Pi(B_n|\boldsymbol{X}) \to 0$ as $n \to +\infty$. We observe that $B_n \subseteq \prod_{i=1}^m \{d(p_i, p_{0,i}) \leq m^{-1/p}M\epsilon_n\}^c$. Moreover, by Theorem D.5 in Ghosal & van der Vaart (2017) and condition 2., the basic testing assumption (4.8) entails that for $i = 1, \dots, m$ there exists test ϕ_{n_i} with error probabilities

$$P_{0,i}^{(n_i)}(\phi_{n_i}) \leq e^{\epsilon_n^2 n_i} \frac{e^{-m^{-2/p}Kn_iM^2\epsilon_n^2}}{1 - e^{-m^{-2/p}Kn_iM^2\epsilon_n^2}};$$
$$\sup_{p \in \mathcal{P}_i \cap \{d(p,p_{0,i}) > m^{-1/p}M\epsilon_n\}} P^{(n_i)}(1 - \phi_{n_i}) \leq e^{-m^{-2/p}Kn_iM^2\epsilon^2}.$$

For $Km^{-2/p}M > 1$, by Lemma 42, there exists a test ϕ that satisfies

$$\left(\prod_{i=1}^{m} P_{0,i}^{n_i}\right)(\phi) \leq 2e^{\epsilon_n^2 n}\frac{e^{-m^{-2/p}KM^2\epsilon_n^2 n}}{1-e^{-m^{-2/p}KM^2\epsilon_n^2 n}};$$

$$\sup_{(p_i)_{i=1}^m \in (\mathcal{P}_1\times\cdots\times\mathcal{P}_m)\cap B_n}\left(\prod_{i=1}^{m} P_i^{(n_i)}\right)(1-\phi) \leq e^{-m^{-2/p}KM^2\epsilon_n^2 n}.$$

Let $A_n = \{\int \prod_{i=1}^m \prod_{j=1}^{n_i} p_i(X_{i,j})p_{0,i}(X_{i,j})^{-1} d\Pi(p_1,\dots,p_m) \geq e^{(C+2m)\bar{\epsilon}_n^2 n}\}$. By Bayes' formula, the posterior probability of B_n is bounded above by

$$\phi + \mathbb{1}_{A_n^c} + e^{(C+2m)\bar{\epsilon}_n^2 n}\int_{B_n}\prod_{i=1}^{m}\prod_{j=1}^{n_i}\frac{p_i(X_{i,j})}{p_{0,i}(X_{i,j})}d\Pi(p_1,\dots,p_m)(1-\phi).$$

The expected value of the first term goes to zero by the previous argument. The one of the second term goes to zero by Lemma 43 with $D = 1$ and condition 1. The expected value of the third term is bounded above by

$$e^{(C+2m)\bar{\epsilon}_n^2 n}\big(\Pi((\mathcal{P}_1\times\cdots\times\mathcal{P}_m)^c) + e^{-m^{-2/p}KM^2\epsilon_n^2 n}\big).$$

Since $\Pi((\mathcal{P}_1\times\cdots\times\mathcal{P}_m)^c) \leq \sum_{i=1}^m \Pi_i(\mathcal{P}_i^c)$, we conclude by condition 3 and by considering $m^{-2/p}KM^2 > C + 2m$. □

4.6.2 Proof of Lemma 42

First of all we prove that ϕ is a test, i.e. a measurable function between 0 and 1. Sum and product of measurable functions are indeed measurable. For every $x_i \in \mathbb{X}_i$, $\phi_i(x_i) \in [0,1]$. We may assume that there exists independent events $\{A_i\}_{i=1}^m$ and a probability measure $\mathbb{P}$ such that $\mathbb{P}(A_i) = \phi_i(x_i)$ for $i = 1,\dots,m$. Then $\phi_n(x_1,\dots,x_m) = \mathbb{P}(\cup_{i=1}^m A_i)$, which is clearly between 0 and 1. To prove condition 1, we reason in a similar way. We observe that $P_{0,i}(\phi_i) \in [0,1]$ and consider independent events $\{A_i\}_{i=1}^m$ such that $\mathbb{P}(A_i) = P_{0,i}(\phi_i)$. Then $\mathbb{P}(\cup_{i=1}^m A_i) \leq \sum_{i=1}^m \mathbb{P}(A_i) \leq m\max(\mathbb{P}(A_1),\dots,\mathbb{P}(A_m)) \leq me^{-a}$. Finally, for condition 3 we consider independent events $\{A_i\}_{i=1}^m$ such that $\mathbb{P}(A_i) = P_i(\phi_i)$. Then $(\prod_{i=1}^m P_i)(1-\phi) = \mathbb{P}((\cup_{i=1}^m A_i)^c) = \mathbb{P}(\cap_{i=1}^m A_i^c)$. Since $(p_i)_{i=1}^m \in \mathcal{U}_0^c$, there exists $\bar{i} \in \{1,\dots,m\}$ such that $p_{\bar{i}} \in \mathcal{U}_{0,\bar{i}}$. Then $\mathbb{P}(\cap_{i=1}^m A_i^c) \leq \mathbb{P}(A_{\bar{i}}^c) < e^{-a_{\bar{i}}}$. We conclude by observing that $a_{\bar{i}} \geq a$.

4.6.3 Proof of Lemma 43

Define $d\Pi_{0,\epsilon}(p_1,\dots,p_m) \propto \mathbb{1}_{\mathcal{V}_{0,\epsilon,n}}(p_1,\dots,p_m)\, d\Pi(p_1,\dots,p_m)$ the restriction of Π to $\mathcal{V}_{0,\epsilon,n}$. The logarithm of the left side of (4.10) is bounded from below by

$$\log(\Pi(\mathcal{V}_{0,\epsilon,n}) + \log\left(\int\prod_{i=1}^{m}\prod_{j=1}^{n_i}\frac{p_i}{p_{0,i}}(X_{i,j})\, d\Pi_{0,\epsilon}(p_1,\dots,p_m)\right).$$

Thus by Jensen's inequality the probability of the complement of the event in (4.10) is smaller or equal to the probability of

$$\int \log\left(\prod_{i=1}^{m}\prod_{j=1}^{n_i}\frac{p_i}{p_{0,i}}(X_{i,j})\right) d\Pi_{0,\epsilon}(p_1,\ldots,p_m) \leq -m(D+1)\epsilon^2 n.$$

Define $Z_i = \int \log(\prod_{j=1}^{n_i} p_i(X_{i,j})p_{0,i}(X_{i,j})^{-1})\, d\Pi_{0,\epsilon}(p_1,\ldots,p_m)$. Then the expected value $\mathbb{E}(Z_i) = -n_i \mathrm{KL}(p_i; p_{0,i}) > \epsilon^2 n$ because of the definition of $\mathcal{V}_{0,\epsilon,n}$. Thus the probability of the complement of the event in (4.10) is smaller or equal to the one of

$$\left\{\sum_{i=1}^{m} Z_i - \mathbb{E}\left(\sum_{i=1}^{m} Z_i\right) \leq -mD\epsilon^2 n\right\},$$

which is bounded from above by $(mD\epsilon^2 n)^{-2}\sum_{i=1}^{m}\mathbb{E}(|Z_i - \mathbb{E}(Z_i)|^2)$ by the triangular inequality and Markov's inequality. The Marcinkiewicz–Zygmund inequality guarantees that $\mathbb{E}(|Z_i - \mathbb{E}(Z_i)|^2) \leq n_i V(p_{0,i}; p_i)$, which is smaller or equal than $\epsilon^2 n$ by definition of $\mathcal{V}_{0,\epsilon,n}$.

4.6.4 Proof of Lemma 44

We approximate the distributions $p_{0,1},\ldots,p_{0,m}$ with convolutions over discrete mixing measure whose atoms are supported on a compact set. Let $a_0 = k_0 \log_-(\epsilon)^{1/r_0}$, where k_0 is a large constant. Define $\tilde{F}_{0,i}$ the restriction of $F_{0,i}$ to $[-a_0, a_0]^d$ for $i = 1,\ldots,m$. Then by Lemma A.3 of Ghosal & van der Vaart (2001), $\|p_{0,i} - p_{\tilde{F}_{0,i},\Sigma_{0,i}}\|_1 \leq 2\epsilon^2$. We argue that for $i = 1,\ldots,m$ there exists a discrete distribution $F_i^* = \sum_{j=1}^{N_i}\omega_{i,j}\delta_{z_{i,j}}$ on $[-a_0, a_0]$ with $N_i \lesssim \log_-(\epsilon)^{d+d/r_0}$ such that $\|p_{0,i} - p_{F_i^*,\Sigma_{0,i}}\|_1 \leq \epsilon^2$. To see this, we apply Corollary B1 of Shen et al. (2013) with $\bar{\epsilon}$ satisfying $\epsilon^2 = \bar{\epsilon}\log_-(\bar{\epsilon})^{d/2}$ and we observe that as $\epsilon \to 0$, $\log_-(\epsilon) \asymp \log_-(\bar{\epsilon})$ up to a multiplicative constant. Since $z \mapsto \phi_\Sigma(x - z)$ is Lipschitz continuous with constant equal to the inverse of the smallest eigenvalue of Σ with respect to the L_1–norm, the points $\{z_{i,j} : j = 1,\ldots,N_i\}$ may be chosen ϵ^2–separated, with possible ties across different groups. We define $U_{i,j} \subseteq [-a, a]^d$ to be a neighborhood of $z_{i,j}$ of diameter ϵ^2 such that if $z_{i,j} \neq z_{i',j'}$, then $U_{i,j} \cap U_{i',j'} = \emptyset$, if $z_{i,j} = z_{i',j'}$, then $U_{i,j} = U_{i',j'}$.
By Lemma B1 in Shen et al. (2013), which extends Lemma 5 in Ghosal & van der Vaart (2007b) in d dimensions, if $\sum_{j=1}^{N_i}|F_i(U_{i,j}) - \omega_{i,j}| \leq \epsilon^2$, then $\|p_{F_i^*,\Sigma_{0,i}} - p_{F_i,\Sigma_{0,i}}\|_1 \lesssim \epsilon^2$. Moreover, if $\|\Sigma_i - \Sigma_{0,i}\| \leq \epsilon$ then $d_H(p_{F_i,\Sigma_{0,i}}, p_{F_i,\Sigma_i}) \leq \epsilon$. Since $d_H^2 \leq \|\cdot\|_1$, we easily conclude.

4.6.5 Proof of Proposition 45

We first recall a known property of the Dirichlet distribution (Lemma 49) and prove an upper bound for the mixed moments of the Pitman–Yor process (Lemma 50). Lemma 49 can be easily deduced by the proof of Lemma 6.1 in Ghosal et al. (2000). The proof

heavily relies on the density of the Dirichlet distribution being bounded from below whenever the parameters are smaller than one. For $N, k \in \mathbb{N}$ and $\gamma \in \mathbb{R}$, let $\{\gamma\}_k$ denote the k–dimensional vector with components equal to γ and let $\mathcal{J}_{N,k} = \{(i,j) : i = 1, \dots N; j = 1, \dots, k\}$.

Lemma 49. *Let $(X_1, \dots, X_k)$ be a random object on $\mathbb{S}^{k-1}$ and let $(u_1, \dots, u_k) \in \mathbb{S}^{k-1}$. Without loss of generality assume $u_k = \max((u_i)_{i=1}^k)$. Then for every $\epsilon \leq k^{-1}$,*

$$\mathbb{P}\left(\sum_{i=1}^{k} |X_i - u_i| \leq 2\epsilon\right) \geq \mathbb{P}(|X_i - u_i| \leq \epsilon^2 \text{ for } i = 1, \dots, k-1)$$

In particular, let $(Y_{1,1}, \dots, Y_{1,k}, \dots, Y_{N,1}, \dots, Y_{N,k}) \sim \mathrm{Dir}(\{\gamma_1\}_\mathrm{k}, \dots, \{\gamma_\mathrm{N}\}_\mathrm{k}$ with $\gamma_i \leq 1$. Then for every $\epsilon \leq (kN)^{-1}$ and $\{u_{i,j}\}$ s.t. $\sum_{(i,j)\in\mathcal{I}_{N,k}} u_{i,j} = 1$,

$$\mathbb{P}\left(\sum_{(i,j)\in\mathcal{J}_{k,N}} |Y_{i,j} - u_{i,j}| \leq 2\epsilon\right) \geq \Gamma\left(k\sum_{i=1}^{N}\gamma_i\right)\epsilon^{2(kN-1)}\prod_{i=1}^{N}\gamma_i^k,$$

where Γ indicates the gamma function.

Lemma 50 yields an upper bound on the mixed moments of the Pitman–Yor process. The proof relies on the relationship between the Pitman–Yor process and the stable completely random measure, together with some convenient tools for evaluating the mixed moments of normalized random measures with independent increments, as first developed in James et al. (2006). For this reason in model (4.1) we focused on $\theta^* > 0$.

Lemma 50. *Let $\tilde{p} \sim \mathrm{PY}(\alpha, \theta, \mathrm{F})$ with $\alpha, \theta > 0$ and $F \in \mathcal{P}_\mathbb{X}$, where $\mathbb{X}$ is a Polish space. The for any $A_1, \dots, A_k$ disjoint Borel sets on $\mathbb{X}$, and $n_1, \dots, n_k \in \mathbb{N}$,*

$$\mathbb{E}(\tilde{p}(A_1)^{n_1} \cdot \dots \cdot \tilde{p}(A_k)^{n_k}) \geq \frac{\Gamma(\theta+1)}{\alpha\Gamma(\theta/\alpha+1)}\frac{\Gamma(k+\theta/\alpha)}{\Gamma(n_+ + \theta)}\prod_{i=1}^{k}\alpha F(A_i),$$

where $n_+ = n_1 + \dots + n_k$ and Γ is the gamma function.

Proof. Let $\mathbb{P}_{\alpha,F}$ be the law of an α–stable completely random measure with base measure F. We define $\mathbb{P}_{\alpha,\theta,F}$ as an absolutely continuous probability with respect to $\mathbb{P}_{\alpha,F}$ with Radon–Nikodym derivative,

$$\frac{d\mathbb{P}_{\alpha,\theta,F}}{d\mathbb{P}_{\alpha,F}}(m) = \frac{\Gamma(\theta+1)}{\Gamma(\theta/\alpha+1)}m^{-\theta}(\mathbb{X}).$$

As shown in Pitman & Yor (1997), the Pitman–Yor process $\mathrm{PY}(\alpha, \theta, F)$ may be obtained by normalizing a random measure $\tilde{\mu} \sim \mathbb{P}_{\alpha,\theta,F}$, i.e.

$$\frac{\mu(\cdot)}{\mu(\mathbb{X})} \sim \mathrm{PY}(\alpha, \theta, F).$$

This relationship between the Pitman–Yor process and the stable completely random measure may be conveniently used to derive the mixed moments, as shown in Canale et al. (2017). In particular, $\mathbb{E}(\tilde{p}(A_1)^{n_1} \cdot \cdots \cdot \tilde{p}(A_k)^{n_k})$ is equal to

$$\frac{\Gamma(\theta+1)}{\Gamma(\theta/\alpha+1)} \frac{1}{\Gamma(n_+ + \theta)} \int_0^{+\infty} u^{n_+ + \theta - 1} e^{-u^\alpha} \prod_{i=1}^{k} \sum_{\ell=1}^{n_i} F(A_i)^\ell \xi_{n_i,\ell}(u)\, du, \tag{4.11}$$

where $\xi_{n,\ell}$ is defined as

$$\xi_{n,\ell}(u) = \frac{\alpha^\ell}{u^{n-\ell\alpha}\ell!} \sum_{\boldsymbol{q}} \binom{n}{q_1 \cdots q_\ell} \prod_{r=1}^{\ell} (1-\sigma)_{q_r - 1},$$

where $(\cdot)_q$ indicates the Pochammer function and the sum is over all vectors $\boldsymbol{q} = (q_1, \ldots, q_\ell)$ of positive integers such that $q_1 + \cdots + q_\ell = n$. We observe that $\sum_{\ell=1}^{n_i} F(A_i)^\ell \xi_{n_i,\ell}(u) \geq F(A_i)\xi_{n_i,1}(u) = F(A_i)\alpha u^{\alpha - n_i}$. Thus we can bound (4.11) from above with

$$\frac{\Gamma(\theta+1)}{\Gamma(\theta/\alpha+1)} \frac{1}{\Gamma(n+\theta)} \left(\prod_{i=1}^{k} \alpha F(A_i) \right) \int_0^{+\infty} u^{\theta + k\alpha - 1} e^{-u^\alpha}\, du.$$

We conclude by observing that the integral in the previous expression is equal to $\alpha^{-1}\Gamma(k + \theta/\alpha)$.

□

We now prove Proposition 45. Conditionally on $\tilde{F}$, $F_1, \cdots, F_m$ are independent. Thus,

$$\Pi(B_\epsilon) = \mathbb{E}\left(\prod_{i=1}^{m} \Pi\left(\sum_{j=1}^{N_i} |F_i(U_{i,j}) - \omega_{i,j}| \leq \epsilon^2 \middle| \tilde{F} \right) \right) \prod_{i=1}^{m} \Pi(\|\Sigma_i - \Sigma_{0,i}\| \leq \epsilon).$$

Since G has continuous and positive density on its support and $\Sigma_{0,i}$ belongs to the support of G, $\Pi(\|\Sigma_i - \Sigma_{0,i}\| \leq \epsilon) \geq \epsilon^q$, where q depends on the dimension of the support of G. Next, let $U_{0,i} = \mathbb{R}^d \setminus (\cup_{j=1}^{N_i} U_{i,j})$, so that $F_i(U_{0,i}), \ldots, F_i(U_{N_i,i}) | \tilde{F} \sim \mathrm{Dir}(\theta\tilde{F}(U_{0,i}), \ldots, \theta\tilde{F}(U_{N_i,i}))$. Let $\lceil\theta\rceil = \min\{n \in \mathbb{N} : n \geq \theta\}$ and let $\eta = \theta\lceil\theta\rceil^{-1} \leq 1$. The aggregation properties of the Dirichlet distribution guarantee that, conditionally on $\tilde{F}$, $F_i(U_{i,j}) = \sum_{h=1}^{\lceil\theta\rceil} Y_{i,j,h}$, where

$$((Y_{i,1,h})_{h=1}^{\lceil\theta\rceil}, \ldots, (Y_{i,N_i,h})_{h=1}^{\lceil\theta\rceil}) \sim \mathrm{Dir}(\{\eta\tilde{F}(U_{i,1})\}_{\lceil\theta\rceil}, \ldots, \{\eta\tilde{F}(U_{i,N_i})\}_{\lceil\theta\rceil}),$$

with $\{\gamma\}_k$ denoting the k–dimensional vector with components equal to γ. Define $\omega_{i,0} = 0$. Then $\sum_{j=1}^{N_i} |F_i(U_{i,j}) - \omega_{i,j}| \leq \sum_{j=0}^{N_i} \sum_{h=1}^{\lceil\theta\rceil} |Y_{h,j,i} - \omega_{i,j}\lceil\theta\rceil^{-1}|$. Lemma 49 thus guarantees that

$$\Pi\left(\sum_{j=1}^{N_i} |F_i(U_{i,j}) - \omega_{i,j}| \leq \epsilon^2 \middle| \tilde{F} \right) \geq \Gamma(\theta)(\epsilon/\sqrt{2})^{2(\lceil\theta\rceil(N_i+1)-1)} \eta^{N_i+1} \prod_{j=0}^{N_i} \tilde{F}(U_{i,j})^{\lceil\theta\rceil}.$$

We observe that $\prod_{i=1}^{m}\Gamma(\theta)(\epsilon/\sqrt{2})^{2(\lceil\theta\rceil(N_i+1)-1)}\eta^{N_i+1} \geq e^{-C_1\log_-(\epsilon)^{d+d/r_0+1}}$ for some $C_1 > 0$, since $N_i \lesssim \log_-(\epsilon)^{d+d/r_0}$. Thus it suffices to show that $\mathbb{E}(\prod_{i=1}^{m}\prod_{j=0}^{N_i}\tilde{F}(U_{i,j})^{\lceil\theta\rceil}) \geq e^{C_2\log_-(\epsilon)^{d+d/r_0+1}}$ for some $C_2 > 0$. We indicate by $\{U_h : h = 1,\dots,N\}$ the set of distinct neighborhoods and by $k_h = |\{i : U_h = U_{i,j} \text{ for some } j\}|$ the number of groups containing a copy of U_h, so that $k_1+\cdots+k_h = N_1+\cdots+N_m =: N_+$. Define $U_0 = \mathbb{R}^d\backslash(\cup_{i=1}^{m}\cup_{j=1}^{N_i}U_{i,j})$, so that $(U_h)_{h=0}^{N}$ forms a partition of $\mathbb{R}^d$ and set $k_0 = m$. Since $U_0 \subseteq U_{0,i}$ for $i = 1,\dots,m$,

$$\mathbb{E}\left(\prod_{i=1}^{m}\prod_{j=0}^{N_i}\tilde{F}(U_{i,j})^{\lceil\theta\rceil}\right) \geq \mathbb{E}\left(\prod_{h=0}^{N}\tilde{F}(U_h)^{k_h\lceil\theta\rceil}\right).$$

In order to compute the expected value on the right, we distinguish two ranges for the parameters. We first deal with the case $\alpha^* = 0$, so that $(\tilde{F}(U_h))_{h=1}^{N}$ has a Dirichlet distribution on $\mathbb{R}^d$. Thus, by known properties of the Dirichlet distribution,

$$\mathbb{E}\left(\prod_{h=0}^{N}\tilde{F}(U_h)^{k_h\lceil\theta\rceil}\right) = \frac{\prod_{h=0}^{N}(\theta^*F^*(U_h))^{[k_h\lceil\theta\rceil]}}{(\theta^*)^{[(m+N_+)\lceil\theta\rceil]}}, \tag{4.12}$$

where $[\cdot]$ is the ascending factorial. Denote by $N_0 = \lceil\theta\rceil(m+N_+) \lesssim \log_-(t)^{d+d/r_0}$. Since F^* is continuous and positive, $\theta^*F^*(U_i) \gtrsim \epsilon^2$ for $i = 1,\dots,N$. Moreover, since $b^k \leq b^{[k]} \leq (b+k-1)^k$, for ϵ sufficiently small the right side of (4.12) is greater or equal to a constant multiplied by

$$\left(\frac{\epsilon^2}{\theta^*+N_0-1}\right)^{N_0} \geq \epsilon^{3N_0} \geq e^{-3\log_-(\epsilon)^{d+d/r_0+1}}.$$

When $\alpha^* > 0$, the expression of the mixed moments is available thanks to the relationship between the Pitman–Yor process and the stable completely random measure. By Proposition 50,

$$\mathbb{E}\left(\tilde{F}(U_0)^{m\lceil\theta\rceil}\prod_{h=1}^{N}\tilde{F}(U_h)^{k_h\lceil\theta\rceil}\right) \gtrsim \frac{\Gamma(N+\theta^*/\alpha^*)}{\Gamma(N_0+\theta^*)}\prod_{h=0}^{N}\alpha^*F^*(U_h).$$

As $\epsilon \to 0$, $N \asymp N_0$. Thus, since $\alpha^*F^*(U_i) \gtrsim \epsilon^2$ for $i = 1,\dots,N$, the right side of the previous expression is bounded from above by $\epsilon^{2N} = e^{-2\log_-(\epsilon)^{d+d/r_0+1}}$.

4.6.6 Proof of Proposition 46

First of all we show that we can reformulate the Kullback–Leibler variation neighborhood in terms of the Hellinger distance (Lemma 51). The proofs follow the lines of Proposition 9.14 in Ghosal & van der Vaart (2017).

Lemma 51. *Let $p_1 = p_{F_1,\Sigma_1}$ be a supersmooth density function as defined in (13) and let $p_2 = p_{F_2,\Sigma_2}$ be a normal mixture density as defined in (4.3). Then for sufficiently small $\epsilon > 0$,*

$$\{KL(p_1;p_2) \lesssim \epsilon^2, V(p_1;p_2) \lesssim \epsilon^2\} \supseteq \{d_H(p_1,p_2)\log(d_H(p_1,p_2)) \lesssim \epsilon\}$$

Proof. By Lemma B.2 of Ghosal & van der Vaart (2017), which builds on Lemma B2 of Shen et al. (2013), both $\mathrm{KL}(p_1;p_2)$ and $\mathrm{V}(p_1;p_2)$ are bounded from above by a multiple of $d_H(p_1,p_2)^2\log(d_H(p_1,p_2))^2$ if, for sufficiently small δ, $P_1((p_1p_2^{-1})^\delta)$ is bounded. For every F_2 and Σ_2 with minimum eigenvalue $\underline{\sigma}_2$, $p_2(x)$ is greater or equal to

$$\frac{1}{\underline{\sigma}_2^d}\int_{\|z\|\leq a} e^{-\|x-z\|^2/(2\underline{\sigma}_2^2)}dF_2(z) \gtrsim \begin{cases} \underline{\sigma}_2^{-d}e^{2da^2/\underline{\sigma}_2^2}F_2([-a,a]^d), & \|x\|_\infty \leq a \\ \underline{\sigma}_2^{-d}e^{2d\|x\|^2/\underline{\sigma}_2^2}F_2([-a,a]^d), & \|x\|_\infty \geq a \end{cases}$$

Since p_1 is uniformly bounded, there exist constants C_1 and C_2 depending on δ and p_1, such that $P_1((p_1p_2^{-1})^\delta)$ is smaller or equal to the sum of $C_1(2a)^d\underline{\sigma}_2^{\delta d}e^{-\delta 2da^2/\underline{\sigma}_2^2}F_2([-a,a]^d)^{-\delta}$ and

$$C_2\underline{\sigma}_2^{\delta d}F_2([-a,a]^d)^{-\delta}\int_{\|x\|_\infty\geq a} e^{-\delta d\|x\|^2/\underline{\sigma}_2^2}p_1(x)\,dx.$$

Since F_1 has sub–Gaussian tails, also p_1 has sub–Gaussian tails. Thus for δ sufficiently small the integral above is finite. We easily conclude by applying a square root transformation. □

For every $i=1,\dots,m$, $p_1=p_{0,i}$ and $p_2=p_i$ satisfy the hypotheses of Lemma 51. By taking ϵ^2 therein equal to $nn_i^{-1}\bar{\epsilon}_n^2$, we deduce that $\Pi(\mathcal{V}_{0,\bar{\epsilon}_n,n})$ is greater or equal to

$$\Pi\left(d_H(p_{0,i},p_{F_i,\Sigma_i})^2\log(d_H(p_{0,i},p_{F_i,\Sigma_i}))^2 \lesssim \frac{n}{n_i}\bar{\epsilon}_n^2 \text{ for } i=1,\dots,m\right). \tag{4.13}$$

We observe that $nn_i^{-1}\bar{\epsilon}_n^2 \geq n_\vee^{-1}\log(n_\vee)^\gamma$, which goes to zero as $n\to+\infty$. Since $\gamma\geq 2$, the probability in (4.13) is greater or equal to $\Pi(d_H(p_{0,i},p_{F_i,\Sigma_i})\leq n_\vee^{-1/2}$ for $i=1,\dots,m)$. By Lemma 44, $\Pi(\mathcal{V}_{0,\bar{\epsilon}_n,n})\gtrsim\Pi(B_{n_\vee^{-1/2}})$, which by hypothesis is greater than $e^{-C\log_-(n_\vee^{-1/2})^\gamma}=e^{-C'n\bar{\epsilon}_n^2}$, where $C'=C2^{-\gamma}$.

4.6.7 Proof of Proposition 47

In order to prove that $\log\mathscr{N}(\epsilon_n,\mathcal{P}_n,d_H)\leq n\epsilon_n^2$, we show that there exist constants C_1,C_2 not depending on n such that $\log\mathscr{N}(C_1\epsilon_n,\mathcal{P}_n,d_H)\leq C_2n\epsilon_n^2$. Indeed, one then defines $\tilde{\epsilon}_n=\epsilon_n\max(\sqrt{C_1},C_2)$ and obtains the desired bound. By Lemma 9.15 in Ghosal & van der Vaart (2017), there exists a large constant A such that $\log\mathscr{N}(A\epsilon_n,\mathcal{P}_n,d_H)$ is smaller or equal to a constant multiplied by

$$N_n\log\left(\frac{5}{\epsilon_n^2}\right)+dN_n\log\left(\frac{3a_n}{\sigma_n\epsilon_n^2}\right)+d^2\log\left(\frac{5}{\epsilon_n^2}\right)+M_nd^2\log(1+\epsilon_n^2)+d\log(M_n).$$

We show that all summands are bounded from above by $n\epsilon_n^2$ up to a constant. First of all we observe that $n\epsilon_n^2>\log(n)$ for sufficiently large n. Thus the last term is smaller or equal to $d^2n\epsilon_n^2$. The fourth one is easily bounded by $d^2n\epsilon_n^2$. Moreover, $\log(5\epsilon_n^{-2})<\log(n)$ for n large enough. Thus, the third term is bounded by $d^2\epsilon_n^2n$. As for the first term, we observe that $N_n(n\epsilon_n^2)^{-1}=(K\log(n))^{-1}$. Thus $N_n\log(5\epsilon_n^{-2})(n\epsilon_n^2)^{-1}<K^{-1}$ for large n.

Similarly, since $a_n\sigma_n^{-1} \leq n^{\alpha_1^{-1}+(2\alpha_2)^{-1}}$, $\log(3a_n\sigma_n^{-1}\epsilon_n^{-2}) < (\alpha_1^{-1}+(2\alpha_2)^{-1}+1)\log(n)$ for large n. Thus $N_n \log(3a_n\sigma_n^{-1}\epsilon_n^{-2})(n\epsilon_n^2)^{-1} < (\alpha_1^{-1}+(2\alpha_2)^{-1}+1)/K$.
We now prove that for every $C > 0$ there exists $K > 0$ such that $\Pi_i(\mathcal{P}_n^c) \geq e^{-Cn\bar{\epsilon}_n^2}$. We observe that $\Pi_i(\mathcal{P}_n^c) \leq \Pi(F_i \in \mathcal{F}_{N_n,a_n}^c) + \Pi(\Sigma_i \in \mathcal{S}_{\sigma_n,M_n}^c)$ and $\Pi(F_i \in \mathcal{F}_{N_n,a_n}^c) = \mathbb{E}(\Pi(F_i \in \mathcal{F}_{N_n,a_n}^c|\tilde{F}))$. Since $F_i|\tilde{F}$ is distributed as a Dirichlet process, by Proposition 2 in Shen et al. (2013), this is bounded from above by

$$\mathbb{E}\left(\left(\frac{2e\theta\log_-(\epsilon_n)}{N_n}\right)^{N_n} + N_n(1-\tilde{F}([-a_n,a_n]^d)\right),$$

which is equal to $(2e\theta\log_-(\epsilon_n)N_n^{-1})^{N_n} + N_n(1-F^*([-a_n,a_n]^d)$. On the other hand, $G(\mathcal{S}_{\sigma_n,M_n}^c) \leq G(\mathrm{eig}_1 \geq \sigma_n^2(1+\epsilon_n^2)) + G(\mathrm{eig}_d \leq \sigma_n^2)$. Putting these together, $\Pi_i(\mathcal{P}_n^c)$ is bounded from above by

$$\left(\frac{2e\theta\log_-(\epsilon_n)}{N_n}\right)^{N_n} + N_n e^{-C_1 a_n^{a_1}} + b_2 e^{-C_2/\sigma_n^{2\alpha_2}} + b_3\sigma_n^{-2\alpha_3}(1+\epsilon_n^2)^{-\alpha_3 M_n}.$$

We show that that for $C > 0$ arbitrarily large, all summands are bounded from above by $e^{-Cn\bar{\epsilon}_n}$. The second and third summand are easily bounded by $e^{-C'n\epsilon_n^2} \leq e^{-KC'n\bar{\epsilon}_n^2}$, for some constant C'. The last summand is bounded by $e^{-2^{-1}\alpha_3 n\epsilon_n^2} \leq e^{-K2^{-1}\alpha_3 n\bar{\epsilon}_n^2}$, by using $1+x \leq e^x$. As for the first summand, we first observe that, for large n, $(2e\theta\log_-(\epsilon_n))^{-1}N_n \geq (n\bar{\epsilon}_n^2)^{1-\delta}$ for $\delta > 0$. Thus for n sufficiently large,

$$\left(\frac{2e\theta\log_-(\epsilon_n)}{N_n}\right)^{N_n} \leq e^{-Kn\bar{\epsilon}_n^2\frac{\log((n\bar{\epsilon}_n^2)^{1-\delta})}{\log(n\bar{\epsilon}_n^2)}} = e^{-K(1-\delta)n\bar{\epsilon}_n^2}$$

By taking K large enough we thus derive the desired upper bound.

Appendix A

Wasserstein distances

If the notion of topology formalizes the concept of convergence, the one of distance allows for its quantification. Since in this thesis we are primarily interested in convergence results for probability laws, we make use of many different distances that, depending on the setting, may capture different properties of the underlying space of probabilities. Among these, the Wasserstein distance plays a predominant role in Chapter 2 and Chapter 3. In this appendix we review some of its defining properties that we repeatedly use throughout the thesis. For a complete reference on the subject we refer to Villani (2008).
Let $\mathbb{X}$ be a Polish with respect to the metric $d_{\mathbb{X}}$ with corresponding Borel σ-algebra. For any pair π_1, π_2 of probability measures on $(\mathbb{X}, d_{\mathbb{X}})$, we indicate by $C(\pi_1, \pi_2)$ the Fréchet class of π_1 and π_2, i.e. the set of distributions on the product space $\mathbb{X}^2$ whose marginal distributions coincide with π_1 and π_2 respectively. If Z_1 and Z_2 are dependent random variables on $\mathbb{X}$ such that their joint law $\mathcal{L}(Z_1, Z_2) \in C(\pi_1, \pi_2)$, we write $(Z_1, Z_2) \in C(\pi_1, \pi_2)$.

Definition 14. The Wasserstein distance of order $p \in [1, +\infty)$ between π_1 and π_2 is

$$\mathcal{W}_{p,d_{\mathbb{X}}}(\pi_1, \pi_2) = \inf_{(Z_1,Z_2)\in C(\pi_1,\pi_2)} \left\{\mathbb{E}(d_{\mathbb{X}}(Z_1, Z_2)^p)\right\}^{\frac{1}{p}}.$$

Let $(\Omega, \Sigma, \mathcal{P})$ be a probability space. By extension, we refer to the Wasserstein distance between two random elements $X_1, X_2 : \Omega \to \mathbb{X}$ as the Wasserstein distance between their laws, i.e. $\mathcal{W}_{p,d}(X_1, X_2) = \mathcal{W}_{p,d}(\mathcal{L}(X_1), \mathcal{L}(X_2))$. An element of $C(X_1, X_2) = C(\mathcal{L}(X_1), \mathcal{L}(X_2))$ is referred to as a coupling between X_1 and X_2.

In this work we will focus on Wasserstein distances on $\mathbb{R}^d$ with respect to the Euclidean distance $\|\cdot\|_d$. In particular we will mainly deal with the Wasserstein distance of order 1 on $\mathbb{R}$ ($\mathcal{W}_1$) and the Wasserstein distance of order 2 on $\mathbb{R}^2$ ($\mathcal{W}_2$), which are arguably the most common choices in the literature. In the main text we denote both these choices by $\mathcal{W}$, hoping that the context clarifies which one we are referring to.

Proposition 52. *The Wasserstein distances on $\mathbb{R}^d$ with respect to the Euclidean distance $\|\cdot\|_d$ enjoy the following properties:*

1. *Let $X, Y : \Omega \to \mathbb{R}$ be random variables. Then,*

$$|\mathbb{E}(X) - \mathbb{E}(Y)| \leq \mathcal{W}_1(X, Y) \leq \mathbb{E}(|X|) + \mathbb{E}(|Y|). \tag{A.1}$$

2. *Let F_X denote the distribution function of X. Then (Dall'Aglio, 1956),*

$$\mathcal{W}_1(X, Y) = \int_{-\infty}^{+\infty} |F_X(u) - F_Y(u)|\, du. \tag{A.2}$$

3. *Let $(\boldsymbol{X}^1, \ldots, \boldsymbol{X}^n)$ and $(\boldsymbol{Y}^1, \cdots, \boldsymbol{Y}^n)$ be n-uples of independent random vectors on $\mathbb{R}^d$, for $d \in \mathbb{N}^+$. Then by Lemma 8.6 in Bickel & Freedman (1981), for $p \geq 1$,*

$$\mathcal{W}_{p,\|\cdot\|_d}(X_1 + \cdots + X_n, Y_1 + \cdots + Y_n) \leq \sum_{i=1}^{n} \mathcal{W}_{p,\|\cdot\|_d}(X_i, Y_i). \tag{A.3}$$

   In particular, (A.3) holds for both $\mathcal{W}_1$ and $\mathcal{W}_2$.

4. *Let $d \in \mathbb{N}^+$ and let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random vectors on $(\mathbb{R}^d, \|\cdot\|_d)$ with finite second moment. Then by Lemma 8.8 in Bickel & Freedman (1981),*

$$\mathcal{W}_2(\boldsymbol{X}, \boldsymbol{Y})^2 = \mathcal{W}_2(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}), \boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y}))^2 + \|\mathbb{E}(\boldsymbol{X}) - \mathbb{E}(\boldsymbol{Y})\|^2. \tag{A.4}$$

5. *Let P_1, P_2, Q_1, Q_2 be probability measures on $\mathbb{R}^d$, for $d \in \mathbb{N}^+$. Then for every $\alpha \in [0, 1]$,*

$$\mathcal{W}_2(\alpha\, P_1 + (1-\alpha)\, P_2, \alpha\, Q_1 + (1-\alpha)\, Q_2) \leq \\ \alpha\, \mathcal{W}_2(P_1, Q_1) + (1-\alpha)\, \mathcal{W}_2(P_2, Q_2). \tag{A.5}$$

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random elements in $\mathbb{R}^d$. A coupling $(\boldsymbol{Z}_X, \boldsymbol{Z}_Y) \in C(\boldsymbol{X}, \boldsymbol{Y})$ is said to be *optimal* if $\mathcal{W}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}(\|\boldsymbol{Z}_X - \boldsymbol{Z}_Y\|^2)^{\frac{1}{2}}$. If an optimal coupling satisfies $\boldsymbol{Z}_X = \phi(\boldsymbol{Z}_Y)$ almost surely for some measurable function ϕ, we refer to ϕ as an optimal (transport) map from $\boldsymbol{X}$ to $\boldsymbol{Y}$. We point out that (A.2) guarantees that optimal maps for the Wasserstein distance on the Euclidean line always exist and are available in an explicit form: if $X, Y \in \mathbb{R}$, $\phi = F_Y^{-1} \circ F_X$ is an optimal transport map from X to Y. On the contrary, in dimension $d > 1$, optimal maps are not guaranteed to exist if $\boldsymbol{X}$ gives non–zero mass to sets of codimension greater or equal to 1. Moreover, even when the existence is established, there is no explicit way to build such maps, except few particular cases; for more details see Villani (2008). Fortunately, there are some sufficient criteria to establish the optimality of a map, as in Theorem 12 of Rüschendorf (1991), which we here specialize to the case $d = 2$.

Theorem 53. *If $\boldsymbol{X}$ is a random object on $\mathbb{R}^2$ and $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ is continuously differentiable, then $(\boldsymbol{X}, \phi(\boldsymbol{X}))$ is an optimal coupling with respect to the 2–Wasserstein distance if and only if the following hold:*

1. *ϕ is monotone, i.e. $\langle \boldsymbol{x} - \boldsymbol{y}, \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) \rangle \geq 0$ for every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2$, where $\langle \cdot \rangle$ indicates the standard scalar product on $\mathbb{R}^2$;*

2. *The matrix $D\phi = \left(\frac{\partial \phi_i}{\partial x_j}\right)_{i,j}$ is symmetric.*

Bibliography

Al Masry, Z., Mercier, S. & Verdier, G. (2017). Approximate simulation techniques and distribution of an extended gamma process. *Methodology and Computing in Applied Probability* **19**, 213–235. 18, 31, 32, 33

Arbel, J., De Blasi, P. & Prünster, I. (2019). Stochastic approximations to the Pitman-Yor process. *Bayesian Analysis* **15**, 1303–1356. 10, 17

Argiento, R., Bianchini, I. & Guglielmi, A. (2016). Posterior sampling from ε-approximation of normalized completely random measure mixtures. *Electronic Journal of Statistics* **10**, 3516–3547. 10

Bacallado, S., Diaconis, P. & Holmes, S. (2015). de Finetti priors using Markov chain Monte Carlo computations. *Statistics and computing* **25**, 797–808. 43, 58

Bickel, P. J. & Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics* **9**, 1196–1217. 17, 65, 99

Blackwell, D. & MacQueen, J. B. (1973). Ferguson distributions via Polya urn schemes. *The Annals of Statistics* **1**, 353–355. 5

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research* **3**, 993–1022. 80

Bondesson, L. (1982). On simulation from infinitely divisible distributions. *Advances in Applied Probability* **14**, 855–869. 31

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability* **31**, 929–953. 8

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I. & Rodrguez, A. (2019a). Latent nested nonparametric priors (with discussion). *Bayesian Analysis* **14**, 1303–1356. 12, 80

Camerlenghi, F., Lijoi, A., Orbanz, P. & Prünster, I. (2019b). Distribution theory for hierarchical processes. *The Annals of Statistics* **47**, 67–92. 12, 80, 90

Camerlenghi, F., Lijoi, A. & Prünster, I. (2020). Survival analysis via hierarchically dependent mixture hazards. *The Annals of Statistics. Forthcoming* . 12, 45, 64

Campbell, T., Huggins, J. H., How, J. P. & Broderick, T. (2019). Truncated random measures. *Bernoulli* **25**, 1256–1288. 10, 17

Canale, A. & De Blasi, P. (2017). Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli* **23**, 379–404. 81

Canale, A., Lijoi, A., Nipoti, B. & Prünster, I. (2017). On the Pitman-Yor process with spike and slab base measure. *Biometrika* **104**, 681–697. 94

Catalano, M., Lijoi, A. & Prünster, I. (2020). Approximation of Bayesian models for time-to-event data. *Electron. J. Statist.* **14**, 3366–3395. xi

Catalano, M., Lijoi, A. & Prünster, I. (2020+). Measuring dependence in the Wasserstein distance for Bayesian nonparametric models. *Under revision* . xi

Chen, J. (1995). Optimal rate of convergence for finite mixture models. *The Annals of Statistics* **23**, 221–233. 17

Cifarelli, D. M., Dolera, E. & Regazzini, E. (2016). Frequentistic approximations to Bayesian prevision of exchangeable random elements. *International Journal of Approximate Reasoning* **78**. 17

Cifarelli, D. M. & Regazzini, E. (1978). Nonparametric statistical problems under partial exchangeability: The role of associative means. Tech. rep. 11, 43

Cifarelli, D. M. & Regazzini, E. (2017). On the centennial anniversary of Gini's theory of statistical relations. *Metron* **75**, 227–242. 17

Cont, R. & Tankov, P. (2004). *Financial Modeling with Jump Processes*. Chapman and Hall/CRC. 14, 61

Cowans, P. J. (2004). Information retrieval using hierarchical Dirichlet processes. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 79

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13. USA: Curran Associates Inc. 18, 19, 44

Daley, D. & Vere-Jones, D. (2002). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Probability and Its Applications. Springer. 7, 18, 46

Daley, D. & Vere-Jones, D. (2007). *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. Probability and Its Applications. Springer New York. 19

Dall'Aglio, G. (1956). Sugli estremi dei momenti delle funzioni di ripartizione doppia. *Annali della Scuola Normale Superiore di Pisa - Classe di Scienze* **10**, 35–74. 99

De Blasi, P., Peccati, G. & Prünster, I. (2009). Asymptotics for posterior hazards. *The Annals of Statistics* **37**, 1906–1945. 28

de Finetti, B. (1937). La prévision : ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré* **7**, 1–68. 3

de Finetti, B. (1938). Sur la condition d'équivalence partielle. *Actual. Sci. Ind.* . 10, 42, 43, 58

De Iorio, M., Johnson, W. O., Müller, P. & Rosner, G. L. (2009). Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics* **65**, 762–771. 16

Doksum, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability* **2**, 183–201. 9, 23, 43, 63

Donnet, S., Rivoirard, V., Rousseau, J. & Scricciolo, C. (2018). Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet process mixtures. *Bernoulli* **24**, 231–256. 17

Dudley, R. (1976). *Probabilities and metrics: convergence of laws on metric spaces, with a view to statistical testing*. Lecture notes series. Aarhus Universitet, Matematisk Institut. 17

Dykstra, R. L. & Laud, P. (1981). A Bayesian nonparametric approach to reliability. *The Annals of Statistics* **9**, 356–367. 9, 16, 23, 27, 28, 31, 35, 43, 45, 64

Elliott, L. T., De Iorio, M., Favaro, S., Adhikari, K. & Teh, Y. W. (2019). Modeling population structure under hierarchical Dirichlet processes. *Bayesian Anal.* **14**, 313–339. 80

Epifani, I. & Lijoi, A. (2010). Nonparametric priors for vectors of survival functions. *Statistica Sinica* , 1455–1484. 14, 46, 61

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230. viii, 3, 4, 6, 8

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics*, M. H. Rizvi, J. S. Rustagi & D. Siegmund, eds. Academic Press, pp. 287 – 302. 9, 81, 82

Ferguson, T. S. & Klass, M. J. (1972). A representation of independent increment processes without Gaussian components. *The Annals of Mathematical Statistics* **43**, 1634–1643. 31

Flamary, R. & Courty, N. (2017). POT Python Optimal Transport library. 22, 51, 57

Gairing, J., Högele, M., Kosenkova, T. & Kulik, A. (2015). Coupling distances between Lévy measures and applications to noise sensitivity of SDE. *Stochastics and Dynamics* **15**, 1550009. 21

Gao, F. & van der Vaart, A. (2016). Posterior contraction rates for deconvolution of Dirichlet-Laplace mixtures. *Electronic Journal of Statistics* **10**, 608–627. 17

Geller, M. & W. Ng, E. (1969). A table of integrals of exponential integral. *Journal of Research of the National Bureau of Standards, Section B: Mathematical Sciences* **73B**. 39

Ghosal, S., Ghosh, J. K. & Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics* **27**, 143–158. 81

Ghosal, S., Ghosh, J. K. & van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics* **28**, 500–531. 92

Ghosal, S. & Van Der Vaart, A. (2007a). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics* **35**, 192–223. 81

Ghosal, S. & van der Vaart, A. (2007b). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics* **35**, 697–723. 92

Ghosal, S. & van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, vol. 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. 16, 23, 43, 86, 90, 95, 96

Ghosal, S. & van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics* **29**, 1233–1263. 81, 92

Gini, C. (1914). Di una misura delle relazioni tra le graduatorie di due caratteri. *Saggi monografici del Comune di Roma, Tip. Cecchini* . 17, 44

Griffin, J. & Leisen, F. (2018). Modelling and computation using NCoRM mixtures for density regression. *Bayesian Analysis* **13**, 897–916. 13

Griffin, J. E. & Leisen, F. (2017). Compound random measures and their use in Bayesian non-parametrics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 525–545. 13, 46, 59

Griffiths, R. & Milne, K. R. (1978). A class of bivariate Poisson processes. *Journal of Multivariate Analysis* **8**, 380–395. 13, 46, 62

Griffiths, T., Canini, K., Sanborn, A. & Navarro, D. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process . 80

Haines, T. S. & Xiang, T. (2011). Delta-dual hierarchical Dirichlet processes: A pragmatic abnormal behaviour detector. In *2011 International Conference on Computer Vision*. IEEE. 80

Hanson, T. E., Jara, A. & Zhao, L. (2012). A Bayesian semiparametric temporally-stratified proportional hazards model with spatial frailties. *Bayesian Analysis* **7**, 147–188. 16

Heinrich, P. & Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics* **46**, 2844–2870. 17

Hjort, N., Holmes, C., Müller, P. & Walker, S. (2010). *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. 16, 43

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics* **18**, 1259–1294. 9, 23, 63

Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173. 10, 17, 43

Ishwaran, H. & James, L. F. (2004). Computational methods for multiplicative intensity models using weighted gamma processes. *Journal of the American Statistical Association* **99**, 175–190. 10, 31

James, L. F. (2003). Bayesian calculus for gamma processes with applications to semiparametric intensity models. *Sankhy: The Indian Journal of Statistics (2003-2007)* **65**, 179–206. 28

James, L. F. (2005). Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *The Annals of Statistics* **33**, 1771–1799. 9, 16, 24, 31, 32, 43, 45, 64

James, L. F., Lijoi, A. & Prünster, I. (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scandinavian Journal of Statistics* **33**, 105–120. 9, 93

James, L. F., Lijoi, A. & Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics* **36**, 76–97. 9

Kallenberg, O. (2017). *Random Measures, Theory and Applications*. Probability Theory and Stochastic Modelling. Springer International Publishing. 46

Kallsen, J. & Tankov, P. (2006). Characterization of dependence of multidimensional Lévy processes using Lévy copulas. *Journal of Multivariate Analysis* **97**, 1551 – 1572. 56, 61

Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics* **21**, 59–78. 7, 43, 46

Kingman, J. F. C. (1975). Random discrete distribution. *Journal of the Royal Statistical Society. Series B. Methodological* **37**, 1–22. With a discussion by S. J. Taylor, A. G. Hawkes, A. M. Walker, D. R. Cox, A. F. M. Smith, B. M. Hill, P. J. Burville, T. Leonard and a reply by the author. 8

Knott, M. & Smith, C. S. (1984). On the optimal mapping of distributions. *Journal of Optimization Theory and Applications* **43**, 39–49. 45, 48

Laud, P. W., Smith, A. F. M. & Damien, P. (1996). Monte Carlo methods for approximating a posterior hazard rate process. *Statistics and Computing* **6**, 77–83. 27, 31

Le Cam, L. (1986). *Asymptotic methods in statistical decision theory.* Springer. 86

Leisen, F. & Lijoi, A. (2011). Vectors of two-parameter PoissonDirichlet processes. *Journal of Multivariate Analysis* **102**, 482 – 495. 14, 46, 61

Leisen, F., Lijoi, A. & Spanó, D. (2013). A vector of Dirichlet processes. *Electronic Journal of Statistics* **7**, 62–90. 13, 60

Lijoi, A., Mena, R. H. & Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association* **100**, 1278–1291. 8

Lijoi, A. & Nipoti, B. (2014). A class of hazard rate mixtures for combining survival data from different experiments. *Journal of the American Statistical Association* **109**, 802–814. 13, 17, 45, 46, 62, 64

Lijoi, A., Nipoti, B. & Prünster, I. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20**, 1260–1291. 13, 46, 62

Lijoi, A. & Prünster, I. (2010). *Models beyond the Dirichlet process.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, p. 80136. 5, 7, 8, 43, 44

Lo, A. & Weng, C.-S. (1989). On a class of Bayesian nonparametric estimates: II. Hazard rate estimates. *Annals of the Institute of Statistical Mathematics* **41**, 227–245. 9, 16, 28, 45, 64

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics* , 351–357. 9, 81

MacEachern, S. N. (1999). Dependent nonparametric processes. *in ASA Proceedings of the Section on Bayesian Statistical Science* , Alexandria, VA: American Statistical Association. 11, 43

MacEachern, S. N. (2000). Dependent Dirichlet processes. *Technical Report* , Ohio State University. 11, 43

Mallows, C. L. (1972). A note on asymptotic joint normality. *The Annals of Mathematical Statistics* **43**, 508–515. 17

Mariucci, E. & Reiss, M. (2018). Wasserstein and total variation distance between marginals of Lévy processes. *Electronic Journal of Statistics* **12**, 2482–2514. 18, 37, 53

Mijoule, G., Peccati, G. & Swan, Y. (2016). On the rate of convergence in de Finetti's representation theorem. *ALEA Lat. Am. J. Probab. Math. Stat.* **13**, 1165–1187. 17

Müller, P., Quintana, F. & Page, G. (2018). Nonparametric bayesian inference in applications (with discussion). *Stat. Methods Appl.* **27**, 175–251. 43

Müller, P., Quintana, F. & Rosner, G. (2004). A method for combining inference across related nonparametric bayesian models . 13

Müller, P., Quintana, F. A., Jara, A. & Hanson, T. (2015). *Bayesian nonparametric data analysis*. Springer Series in Statistics. Springer, Cham. 23, 43

Nakamura, T., Nagai, T. & Iwahashi, N. (2011). Multimodal categorization by hierarchical Dirichlet process. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 80

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics* **41**, 370–400. 17, 44

Nguyen, X. (2016). Borrowing strength in hierarchical Bayes: Posterior concentration of the Dirichlet base measure. *Bernoulli* **22**, 1535–1571. 44, 80

Nipoti, B., Jara, A. & Guindani, M. (2018). A Bayesian semiparametric partially PH model for clustered time-to-event data. *Scandinavian Journal of Statistics* **45**, 1016–1035. 17

Panaretos, V. M. & Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application* **6**, 405–431. 17, 44

Pennell, M. L. & Dunson, D. B. (2006). Bayesian semiparametric dynamic frailty models for multiple event time data. *Biometrics* **62**, 1044–1052. 16

Perman, M., Pitman, J. & Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields* **92**, 21–39. 6

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability theory and related fields* **102**, 145–158. 5, 6

Pitman, J. (2003). Poisson-Kingman partitions. *Institute of Mathematical Statistics Lecture NotesMonograph Series IMS, Beachwood, OH. Mathematical Reviews* **40**, 134. 8

Pitman, J. (2006). *Combinatorial Stochastic Processes: Ecole d'Été de Probabilités de Saint-Flour XXXII-2002*. Springer. 5

Pitman, J. & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25**, 855–900. 6, 82, 93

Quintana, F. A., Mueller, P., Jara, A. & MacEachern, S. N. (2020). The dependent Dirichlet process and related models. *arXiv preprint arXiv:2007.06129* . 11, 80

Rachev, S. (1985). The Monge-Kantorovich mass transference problem and its stochastic applications. *Theory of Probability & Its Applications* **29**, 647–676. 17, 44

Regazzini, E. (1996). *Impostazione non parametrica di problemi d'inferenza statistica bayesiana*. Consiglio nazionale delle ricerche. 1

Regazzini, E., Lijoi, A. & Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics* **31**, 560–585. Dedicated to the memory of Herbert E. Robbins. 8, 43, 63

Riva Palacio, A. & Leisen, F. (2018). Bayesian nonparametric estimation of survival functions with multiple-samples information. *Electronic Journal of Statistics* **12**, 1330–1357. 13

Riva Palacio, A. & Leisen, F. (2019). Compound vectors of subordinators and their associated positive Lévy copulas. *arXiv 1909.12112* . 46

Rodrguez, A., Dunson, D. B. & Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association* **103**, 1131–1154. 12, 80

Rosiński, J. (2001). *Series representations of Lévy processes from the perspective of point processes*. Boston, MA: Birkhäuser Boston, pp. 401–415. 31

Rüschendorf, L. (1991). Fréchet-bounds and their applications. *Advances in Probability Distributions with Given Marginals* **67**. 45, 48, 65, 99

Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge Studies in Advanced Mathematics. Cambridge University Press. 36, 69

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **4**, 10–26. 81

Scricciolo, C. (2014). Adaptive Bayesian density estimation in l^p -metrics with Pitman-Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Analysis* **9**, 475–520. 90

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650. 6, 43

Shen, W., Tokdar, S. T. & Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100**, 623–640. 81, 89, 90, 92, 96, 97

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B. & Lanckriet, G. R. G. (2012). On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics* **6**, 1550–1599. 18, 19

Srivastava, S., Li, C. & Dunson, D. (2015). Scalable Bayes via barycenter in Wasserstein space. *Journal of Machine Learning Research* **19**. 17

Taniguchi, T., Yoshino, R. & Takano, T. (2018). Multimodal hierarchical Dirichlet process-based active perception by a robot. *Frontiers in Neurorobotics* **12**, 22. 80

Tankov, P. (2003). Dependence structure of spectrally positive multidimensional Lévy processes. *Unpublished manuscript* . 14, 46, 61

Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101**, 1566–1581. 12, 79, 80

Tokdar, S. T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics* , 90–110. 81

Trippa, L. & Favaro, S. (2012). A class of normalized random measures with an exact predictive sampling scheme. *Scandinavian Journal of Statistics* **39**, 444–460. 10

Villani, C. (2008). *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg. 17, 19, 44, 48, 98, 99

Wolpert, R. L. & Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika* **85**, 251–267. 31

Wu, Y. & Ghosal, S. (2010). The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *Journal of Multivariate Analysis* **101**, 2411 – 2419. 81

Xing, E. P., Jordan, M. I. & Sharan, R. (2007). Bayesian haplotype inference via the Dirichlet process. *Journal of Computational Biology* **14**, 267–284. 80

Zhou, H., Hanson, T., Jara, A. & Zhang, J. (2015). Modeling county level breast cancer survival data using a covariate-adjusted frailty proportional hazards model. *The Annals of Applied Statistics* **9**, 43–68. 17